1-1-2017

Insights On Fibrotic Diseases Using Single-Cell Analysis Methods

Azim Munivar
Yale University

“Insights on fibrotic diseases using single-cell analysis methods”

A Thesis Submitted to the
Yale University School of Medicine
in Partial Fulfillment of the Requirements for the Degree of Doctor of Medicine

by

Azim Munivar

2017

Advisor: Naftali Kaminski

Thesis Committee: Naftali Kaminski, Erica Herzog, Hongyu Zhao

Abstract

Idiopathic pulmonary fibrosis is progressive, fatal lung disease with unclear mechanistic etiology and a dearth of treatment options. Transcriptional profiling has served a valuable tool in understanding the underlying perturbations in the lung tissues of patients and disease model systems, however whole tissue profiling obscures both the contribution of individual cells types in the diseased tissue as well as the contribution of non-fibrotic tissue surrounding the diseased tissue. The averaging effect confounds the ability to extract a strong disease signal and understand the cell-of-origin. Single-cell techniques have recently emerged that allow profiling of the transcriptomes of individual cells. In this work, we employ two state-of-the-art single cell RNA-seq techniques to IPF-relevant disease systems to understand cell specific contributions. In one set of experiments, we extracted and dissociated lung tissue from Tgfβ1 induced, as well as bleomycin injured mice systems. Single cells were isolated into individual wells using the Fluidigm C1 Auto Prep Array IFC system and single cell libraries were generated and sequenced. We observed the upregulation of fibroblast specific genes in cells with epithelial cell markers reinforcing theories of epidermal to mesenchymal transition. In another set of experiments, we used a high-throughput, droplet-based system to study the knockdown of FENDRR, a novel long-non coding RNA (lncRNA) implicated in lung fibrosis in normal human lung fibroblasts (NHLFs). Here we observed cell-specific upregulation of genes associated with fibrosis and quiescence, as well as a stochastic effects demonstrating cell-cycling that would have otherwise been indiscernible without single-cell methods. In this work, we also address the

significant challenges in creating robust single cell libraries using both human and mouse tissue. These challenges, shortcomings, and future opportunities for single-cell sequencing are highlighted.

Acknowledgements

Naftali Kaminski

Hongyu Zhao

Erica Herzog

Farida Ahangari

Taylor Adams

Koji Sakamoto

Nachelle Aurelian

Table of Contents

Introduction

1) Idiopathic Pulmonary Fibrosis Background

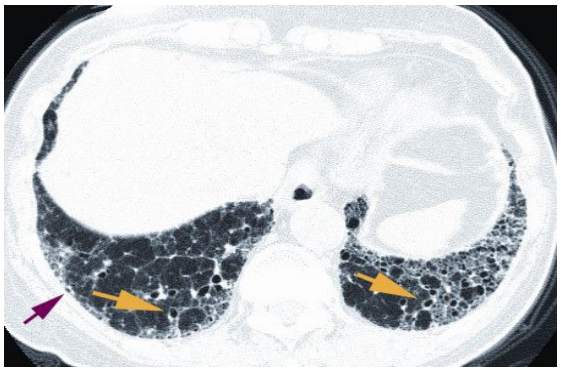
Idiopathic fibrosis is a chronic, progressive, lethal, interstitial lung disease. The disease incidence is conservatively estimated at 3 to 9 cases per 100,000(1), and thus it is relatively rare, but within certain age groups, the disease is more lethal than a diagnosis of lung cancer.

The core process at play in this disease is the accumulation of fibrotic lesions in the lung that replace the normal lung parenchyma(2). This results in both a restriction in respiration due to the increased stiffness within the lung, as well as a loss of diffusion capacity due to both loss of normal lung parenchyma as well as the increased separation between capillaries and the inner compartment of the alveoli. Forced vital capacity (FVC) and diffusion capacity for carbon monoxide (DLCO) are thus the standard staging and progression criteria(3). In the normal progression of the disease, patient lung function declines steadily, and patients are prone to lung infections causing step-function like declines in FVC and DLCO(3, 4). Loss of these physiologic parameters is thus the mechanism of death for these patients. Complications of the disease include pulmonary hypertension, thromboembolic disease, and lung cancer(5). Most of these are thought to be secondary to the fibrotic process, but as disease etiology is unclear, these complications may have contributions from the primary etiology or etiologies.

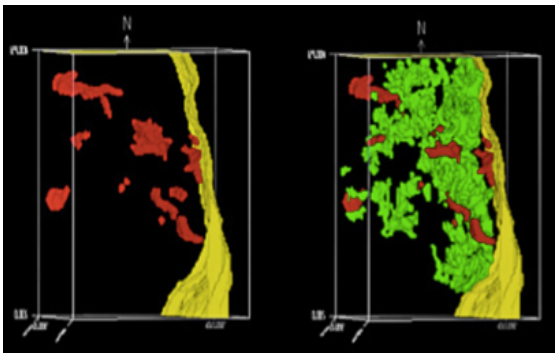
Histopathology

Classically, the histologic diagnosis of IPF is based on the pattern of usual interstitial pneumonia (UIP)(6). While current classification relies on integration of

A



B



C

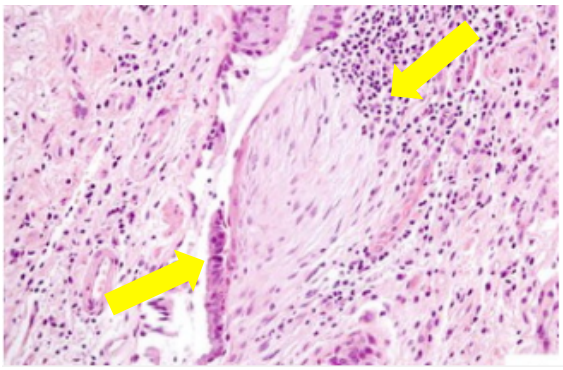


Figure 1: A. High-resolution CT scan of IPF lung demonstrating pervasive honeycomb cysts (yellow arrows) in the subpleural region (red arrow) (Image courtesy of Naftali Kaminski). **B**. Computer reconstruction of the fibrotic reticulum (Pleura – Yellow. Blood vessels – Red. Fibrotic reticulum – Green)

histologic, radiologic and clinical criteria, the underlying pathologic changes are still understood histologically. UIP is marked by fibrotic regions potentially positive for honeycombing in subpleural and paraseptal regions (6) as seen in Figure 1.A and 1.B(7). The location specific differences in the disease are a critical differentiating factor with normal tissue seen adjacent to established disease. Immature hyaluronic acid rich matrix secreted by fibroblastic foci (Figure 1.C) is present near recently generated scars and is bound by areas of significant epithelial cell damage and death(6, 8). This process causes a regeneration response and proliferation of alveolar type II cells(9). Thus there is heterogeneity both from a temporal perspective with active disease in some regions and mature scar in others, as well as spatial differences in the extent of tissue injury seen(6). Notably, there is an typically an absence of necrotic debris and inflammatory cell infiltrate(10).

Pathogenesis

IPF pathogenesis remains unclear but the disease is thought to occur due to recurrent injury to the alveolar epithelium and is perpetuated by an abnormal repair process resulting in fibrosis that extends from the epithelium to the interstitium(8). The injury is thought to alter the basement membrane and permits entry of mesenchymal cells(11) causing a complex interplay of cytokines, basement membrane driven cellular changes, and non-classical inflammation resulting in highly active, contractile fibroblasts(12).

These activated fibroblasts and an emerging population of myofibroblasts organize to form fibrotic foci (Figure 1.A) that are precursors to end-stage fibrosis(7). These foci develop near locations of epithelial injury and basement membrane disruption. The persistence of these foci may be due to inhibition of normal apoptotic processes as well as initiation of senescence. Targeting this senescence process has been shown to reverse fibrosis that accumulates due to oxidative injury in aging mice(13).

Centrality of Tgfβ1

Transforming growth factor beta 1 (Tgfβ1) is a highly conserved, and ubiquitously expressed cytokine important in tissue growth, injury, and repair across many disease categories including cancer, inflammation, and fibrosis(14). It plays a central role in IPF pathogenesis in multiple lines of investigation. It is increased in tissue samples from patients with IPF(15) as well as in tissue in animal models of the disease(16). Animal models also demonstrate that overexpressing active Tgfβ1

leads to lung fibrosis(17) and blocking the Tgfβ1 receptor I (ALK5) (18)and it's down stream signaling pathways significantly reduces or prevents fibrosis (18, 19).

Tgfβ1 signaling occurs through the interaction of active Tgfβ1 with the Tgfβ1 receptors I and II on target cells(20, 21). As Tgfβ1 is synthesized and held in an inactive form in a latent complex, it must be activated. It can be activated though a number of processes including physical changes in parameters like pH; extreme temperature changes; enzymatic cleavage via a number of proteases including plasmin, tryptase, and MMP-2 and -9; as well through interactions with integrins(22). The integrin activation has received special attention as integrins allow for direct communication and integration of extracellular and intracellular signals to the external environment(23). In this case, the avB6 integrin has been implicated in initiating and perpetuating fibrosis in various murine models(24). Moreover, blocking certain integrin signaling, such as avB1 and avB6, has been shown to ameliorate or prevent fibrosis(25, 26).

Activated Tgfβ1 stimulates fibroblasts to transdifferentiate into myofibroblasts and induces epithelial to mesenchymal transition in epithelial cells. Tgfβ1 is also a potent inducer of extracellular matrix deposition(27).

Relevance of single cell methods

Because of the extreme heterogeneity of cell types in the lung and the spatial and temporal variability in disease, tissue level measurements of gene expression are likely to obscure specific signals of disease that provide greater insight on disease etiology or mechanism(28).

The progressive nature of IPF, the complex interplay of cytokines, and the emergence of transdifferentiated cell types points to a key need for single-cell analysis. For example, the close proximity of mesenchymal cells to epithelial cells may contribute to the regulation of cell proliferation and connective tissue synthesis by fibrogenic cytokines released from epithelial cells. The current hypothesis suggests that repeated subclinical injury to the lung injures the alveolar epithelial, the subepithelial, and adjacent endothelial basement membranes(9). This injury permits entry into the alveoli of cells of the mesenchymal lineage, thus resulting in a mixed population even within a rather discrete a well-defined cellular niche.

Moreover, there is known to be heterogeneity in collagen expression from the fibroblast population indicating a potential underlying phenotypic difference. In situ hybridization experiments have demonstrated that fibroblastic foci specifically showed an increase in expression of type I procollagen mRNA which was not observed in other areas(29).

Cell populations also shift within IPF lungs. Microscopic characterization of normal lungs indicate that 8% of cells are AT1, 16% are AT2, and cells in the interstitial space are 37% of the total cell number with a subset of these of being fibroblasts(30). Given the fibrotic changes that occur in the lung, including the development of the histological hallmark fibroblastic foci occurring at sites of putative epithelial injury, there is an expansion of the fibroblast population in IPF lungs, however the source and exact quantification changes in cell number is not well described(31, 32).

Given the importance of the fibroblast in the proliferation of IPF, characterizing the phenotypes and underlying expression differences between the normal appearing fibroblasts that reside in the normal regions of the IPF lung relative to both the normal appearing fibroblasts in diseased areas as well as the diseased fibroblasts in fibrotic foci will provide important clue regarding intermediate disease states and disease progression and diversity within an individual.

Cell types of interest in the fibrotic process

Type I and II Alveolar epithelial cells

Alveoli are lined by squamous type 1 alveolar epithelial cells (AT1) and cuboidal type 2 alveolar epithelial cells (AT2)(33). AT1 cells serve as the primary cells across which gas exchange occurs(34). Their flat shape facilitates the rapid diffusion of gas and they are characterized by expression of Hopx, podoplanin (Pdpn/T1alpha), and AGER)(35). AT2 are considered a stem cell that will differentiate into AT1s and clonally expand to replenish AT2s in situations of lung injury(36). Additionally, AT2s serve as secretors of surfactant and are marked by expression of SftpA-C and LysM(35, 37).

In IPF, AT2s proliferate in response to injury but do not normally re-epithelialize the alveolar space(38, 39), potentially due to persisting abnormalities such as fragmentation of alveolar basement membrane components such as hyaluronic acid(40, 41). In fact, both epithelial injury and basement membrane injury appear necessary for the development of interstitial fibrosis. Abnormalities in the alveolar basement membrane provide the impetus for migration of mesenchymal cells into the alveolar spaces, followed by deposition of collagen preventing the

expansion of a collapsed airspace(42). The perpetuation of this process by chronic injury results in a continuing process of fibrosis and remodeling. Recent single cell work in isolated epithelial cells from normal and diseased lung has shown that in disease three subsets of epithelial cells exist all of which show abnormal differentiation. These cells frequently co-express markers of AT1 and AT2 cells, as well as conducting airway specific markers. These cells are not thought to be related to normal undifferentiated lung progenitor cells, and instead likely represent a failure to suppress multilineage differentiation programs(43).

Fibroblasts

Fibroblasts comprise a large component of the stromal population of the lung parenchyma along with myofibroblasts, pericytes, and lipofibroblasts. They primarily secrete and degrade collagen in a tightly regulated process to preserve the normal lung structure. They secrete both type I and type II collagens and serve to break down up to 40% of newly synthesized collagen(44, 45).

In IPF, the population of collagen secreting fibroblasts is thought to be increased especially within fibrotic foci(7, 42, 46). These fibroblasts lie on the alveolar side of the injured airspace as evidenced by detection of residual areas of basal lamina(47). Fibroblasts isolated from lungs with early fibrosis show much greater replicative potential than fibroblasts from normal lungs, or those from lungs with late stage fibrosis(48). Moreover, late stage fibroblasts from IPF lungs also demonstrate an increased senescent associated phenotype as determined by resistance to oxidative stress(49). Thus the phenotype of fibroblasts appears to

progress through two stages, initially characterized by increased proliferation in response to injury, and later by lack of apoptosis and cellular senescence.

Myofibroblasts

Myofibroblasts express both features of fibroblasts as well as smooth muscle cells and are marked by expression of alpha-smooth muscle actin (SMA) expressed by the gene ACTA2(50). Notably, they produce collagen at higher levels than normal fibroblasts(51). In IPF, myofibroblasts emerge from the differentiation of fibroblasts in response to Tgfβ1, as well as other cytokines(40). Epithelial mesenchymal transition, driven by growth factors and transcription factors, is thought to be another source for IPF myofibroblasts(52, 53). Tgfβ1, epidermal growth factor (EGF), hepatocyte growth factor (HGF), and fibroblast growth factor (FGF) can cause epithelial cells to acquire myofibroblast markers including SMA, FSP1 and lose epithelial markers including E-cadherin and ZO-1(53).

Myofibroblasts from IPF lungs also show microRNA abnormalities relative to normal lungs including increased levels of mIR-21 and decreased expression of mIR-29. mIR-21 promotes proliferation and differentiation and mIR-29 inhibits synthesis of fibrogenic proteins(54). As microRNAs regulate global transcriptional process, IPF myofibroblasts represent a population of cells with widespread dysfunctions in protein expression potentially at the post-transcriptional level.

2) Model systems for IPF

Bleomycin induced lung injury model

The bleomycin mouse injury model involves intratracheal administration of a single dose of bleomycin. Bleomycin was previously used to treat cancer but quickly

fell out of favor as it was found to cause pulmonary fibrosis as a side effect of its chemotherapeutic purpose in certain populations. This was found to be due to low levels of the hydrolyzing enzyme for bleomycin, bleomycin hydrolase in the lungs of certain individuals. The injury from bleomycin occurs due to a combination of direct DNA strand breakage, free radical generation, and induction of oxidative stress.(55)

The intratracheal instillation of bleomycin causes injury to alveolar epithelium proceeded by a wave of activated macrophages and neutrophil-rich inflammatory infiltrate that resolves in one to three weeks(56-58). After the initial inflammation is cleared, fibroblast proliferation is noted and significant ECM is deposited. The fibrotic response in this model can be seen biochemically and histologically by day 14 with maximal responses at days 21-28. (57). In this process, myofibroblasts and fibroblast both increase, but while fibroblast number return to base line values after four weeks, myofibroblast quantities are 10-fold higher than in the normal lung(59), leading to potentially persistent production of ECM(60). Another study demonstrated development of two distinct mesenchymal cell populations one located in areas of active fibrosis and positive for SMA, desmin, and procollagen, and the other located in submesothelial areas and expressing only SMA and procollagen(61).

Tgfβ1 overexpression model

As Tgfβ1 is thought to play a major role in the development of IPF, model systems have been developed that overexpress Tgfβ1 in a lung specific fashion in order to understand the specific impact of this model system(17). A triple transgenic mouse was developed to express biologically active Tgfβ1 in a tightly inducible fashion. Induction of Tgfβ1 in this model caused transient apoptosis of epithelial

cells, followed by inflammation rich in mononuclear cells, parenchymal and interstitial fibrosis, myofibroblast hyperplasia, and eventual alveolar septal rupture and honeycombing. An important feature of these mice was the reversibility of the fibrotic process. Withdrawal of the doxycycline stimulus inducing Tgfβ1 production resulted in resolution of fibrosis and normalization of collagen content in the affected lung after only one month. The investigators further characterized the mechanism of fibrosis in this model focusing on the initial event of epithelial apoptosis. Apoptosis was inhibited using a variety of strategies, and was found to be a necessary condition for the Tgfβ1 induced fibrotic response

Other model systems for IPF

Multiple model systems for IPF exist outside of the bleomycin instillation and Tgfβ1 overexpression models described above. Several of these models involve delivery of an injurious chemical or energetic stimulus. Instillation of FITC, a fluorescent molecule, causes fibrosis that is dependent on CCR2 signaling recruiting fibrocytes(62) and production of IL-13(63). The advantage of this model is that fibrosis can be localized with by immunofluorescence imaging for green color of FITC(64) and response is persistent for at least six months(65). Irradiation at a dose of 12-15 Gy can also induce fibrosis, but does so dependent on the mouse strain with C57Bl/6 mice being the most fibrosis prone. The fibrosis is thought to be induced by an interplay of TNF-a and Tgfβ1 signaling depending on the mouse strain(66). Thus the advantage to this model system is that it allows investigations into genetic causes of fibrosis, but requires housing mice for at least the 20-24 weeks it takes for fibrosis to develop(67). Silica instillation is another option for

inducing fibrosis. This system has advantages of also being strain dependent (68) facilitating genetic understanding of silica-induced fibrosis, and as silica is not cleared from the lung, the fibrotic stimulus is persistent(69).

Other transgenic models have been developed to induce overexpression of pro-fibrotic stimuli such as GM-CSF, TNF-a, and IL-1B(70). These have been delivered via AAV allowing for transient expression, but somatic and germline lentiviral-mediated incorporation of fibrotic transgenes has also been demonstrated.

3) FEDNRR, a novel lncRNA plays a role in cell senescence

FENDRR is a lung-specific long non-coding RNA (lncRNA). LncRNA are a type of regulatory RNA that play a prominent role in development, differentiation, and epigenetic modification. Aberrant lncRNA expression has been implicated a number of human diseases, but has not been well characterized in the complex process of pulmonary fibrosis(71). FENDRR's role in cellular senescence in normal human lung fibroblasts was first characterized in our lab. FENDRR was selected based on a high-throughput microarray screen of lincRNA expression across 400 lung samples in patients with COPD and various forms of interstitial lung disease(72). FENDRR was found to be the most downregulated lincRNA in IPF lungs, and was validated using the nCounter assay (results under publication). FENDRR knockdown in NHLF showed a prominent role for induction of cellular senescence as well as myofibroblastic differentiation. Mechanistic studies showed that loss of FENDRR results in epigenetic remodeling through the PRC2 complex and alters methylation on the p16 and GATA6 promoters to activate downstream expression of senescence associated genes(73). Aging has been shown to play a key risk factor in the

development of IPF and FENDRR points to a specific epigenetic mechanism by which cellular senescence is regulated in fibroblasts.

4) Single Cell Methods

C1

The C1 is an instrument designed by Fluidigm corporation to enable automation and reproducibility in single cell genomic experiments(74). The instrument, and associated chips isolate single cells into microliter wells and precisely deliver reagents into those wells. These reagents can be specific to the protocol at hand, and at the moment protocols exist for high-throughput sequencing of cellular mRNA, miRNA, methylation changes, chromatin marks, and other genomic signatures(75).

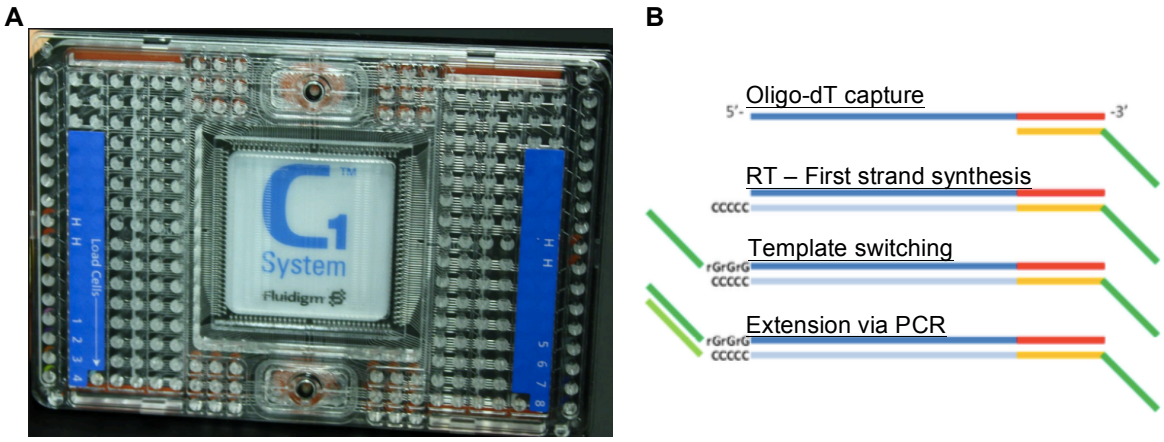


Figure 2: Key features of the C1. A. A typical IFC is shown here. Reagents and cells are loaded at pre-determined locations and reaction products are harvested from specific well locations. **B.** Demonstrates the SMART-seq protocol for amplification of cellular RNA using oligo-dT capture and template switching. Captured RNA (dark blue) with polyA tail (red) is primed for first strand synthesis by an oligo-dT sequence primer (yellow) with a common PCR handle on the 5' end (green). First strand extension adds untemplated dTs. A template switching primer with 3' ribosylated guanosine bases and a 5' common PCR handle is used to create the second strand of DNA using the 'first strand' as a template. The common PCR handles can now be used to amplify the newly derived cDNA.

The instrument operates on a proprietary microfluidics technology that automatically flows a cell suspension through a complex array of gates, detects the presence of a single entering cells through voltage sensitive gates, captures that cell, and diverts the remaining cell suspension to the next well . An example chip is shown in Figure 2.A. The chip can then be physically inspected under a microscope to confirm capture of single cells, or alternatively to note the presence of empty wells or cell doublets. Reagents can also be added to stain the cells to determine cell viability.

For the mRNA-seq system employed to analyze single cells in this paper, the procedure utilizes the SMARTer template switching reaction(76). Briefly, a lysis reagent lyses the cells and stabilizes the liberated RNA, after which an oligo-dT capture and priming method is used to prime first strand synthesis of cDNA (Figure 2.B). The specialized reverse transcriptase adds non-templated dC nucleotides to the terminal 3' end of the nascent strand. A template switching primer with a common PCR tag on the 5' side and dG nucleotides on the 3' side is used to prime this first strand and confer a common PCR tag. PCR is then used to amplify and enrich these properly captured sequences. Enriched cDNA is then moved into a 96-well plate where Illlumina Nextera library preparation occurs. Each sample is given a well-specific barcode by addition of unique PCR primers during this process and then the 96 wells are pooled for final preparation for sequencing and sequencing. The barcodes are then used to demultiplex the sequencing results and data processing pipeline is used to determine the abundance of expressed genes in each cell. Subsequently, low quality cells and wells containing cells known to be doublets

or empty wells are filtered out, reads are aligned to a known genome, and counted using one of several quantification methods.

Drop-Seq

Drop-Seq was published in 2015 and functions by co-encapsulating single cells and barcoded beads into nanoliter droplets(77). At the time it was released it increased throughput 50-100 fold and reduced the per-cell cost single cell library assembly by over 100 times relative to the C1. The clear advantages of this system provided a compelling reason for our lab to move forward with implementing this technology.

The Drop-Seq platform leverages several advances in microfluidics, high throughput sequencing, microparticle chemistry, and barcode-based molecular biology. A reproduced image demonstrating the main feature of the system is shown in Figure 3(77). The system is assembled from commercially available equipment and reagents. Three syringe pumps, holding a cell suspension, barcoded beads, and oil, with the former two in an aqueous suspension, are simultaneously flowed into a microfluidic chip (Figure 4). The beads are synthesized such that each bead is decorated with a millions of barcoded primer sequences, which are identical on a given bead, but distinct from every other bead (Figure 3.B, 3.C). A nozzle at the flow junction of the aqueous solution and oil solution results in surface tension forces that causes formation of droplets at the rate of approximately 80,000 per second (Figure 4.C)(77).

Beads and cells are encapsulated randomly, subject to Poisson probability

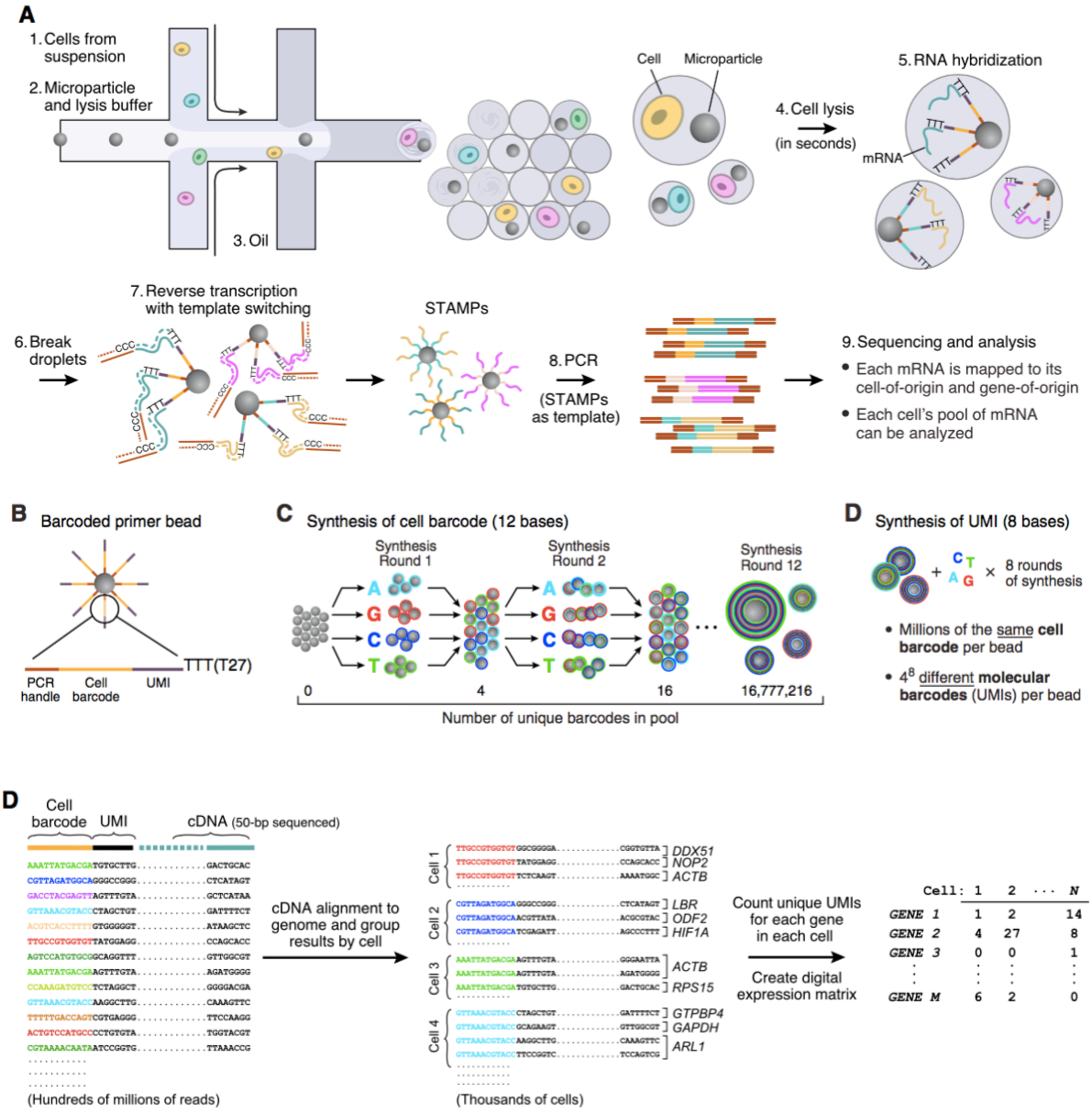


Figure 3: Overview of the Drop-Seq Protocol. A. Cells and beads are combined in a microfluidic device to co-encapsulate beads and cells. Lysed cells are then subject to reverse transcription to allow cellular polyA-RNA to inherit unique barcodes from beads. Libraries are generated from this first strand of DNA and sequenced using paired end sequencing. **B.** Beads are designed with three regions containing a common PCR handle, a cell barcode region, and a UMI. **C.** The diversity of cell barcodes and UMIs is generated using 12 rounds of a split-pool synthesis design for the cell barcode and 8 round of random addition for the UMI. **D.** Paired-end reads are computationally de-multiplexed by grouping reads with the same cell barcode to one cell. Reads are subsequently collapsed by UMI to control for amplification biases. (Figure reproduced with permission of Cell/Elsevier).

distributions (Figure 5). In order to achieve droplets where only one cell and one bead end up in a droplet, a dilute cell and bead suspension must be created so that on average, less than one cell or bead becomes encapsulated into a droplet. Thus, most droplets formed are empty (Figure 4.D) and a small number contain a bead or a cell, and an even smaller number contain both a bead and a cell. This results in only a small fraction of the total number of cells flowed in becoming successfully

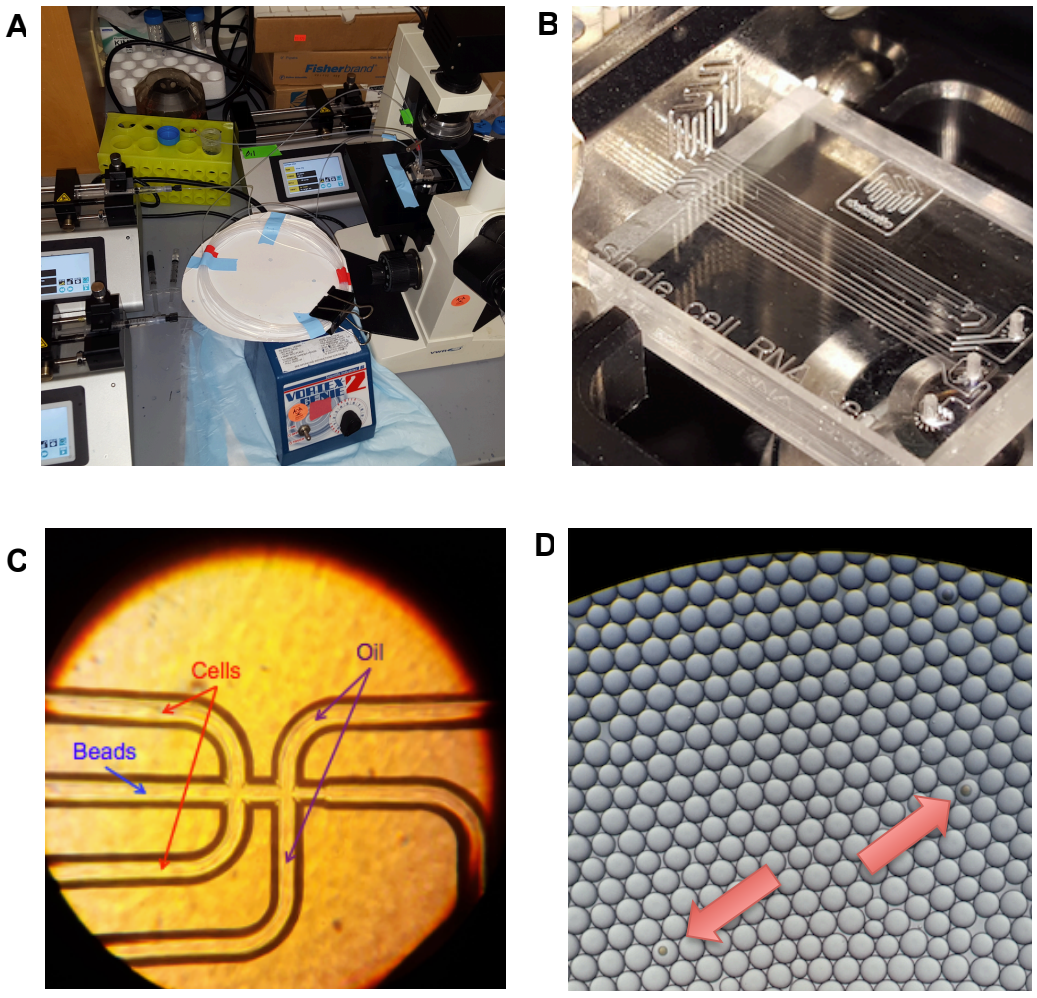


Figure 4: Drop-Seq Setup Photos. A. Overall set up Drop-Seq Rig. Three syringe pumps with oil, cells, and beads feed into a microfluidic chip shown in **B. C.** Microfluidic channels visualized under a microscope at 10x magnification. **D.** Droplets formed under a microscope. Rare droplets receive a bead and a smaller fraction also contain a co-encapsulated cell.

sampled. Despite this limitation, the large number of total cells are sampled thus resulting in a representative view of the whole population(77).

The bead solution contains a lysis reagent which, shortly after mixing with the cell solution, causes breakage of the cell membrane within the encapsulated droplet releasing the contents of the cell including the mRNA. Polyadenylated mRNA binds to a ‘poly-T’ sequence on the beads. The droplets are captured in a Falcon tube as oil-water emulsion and all further processing occurs in bulk for a given sample. In the next step, the droplets are broken releasing the mRNA coated beads (termed STAMPs – “single-cell transcriptomes attached to microparticles”), washed, and incubated with a reverse transcriptase. This droplet breakage step also releases significant ambient RNA from cells that did not associate with a bead as well as beads that did not co-encapsulate with a cell. This necessitates very quick preparation and several washing steps to eliminate this unbound mRNA to minimize the technical noise. During the reverse transcription step, transcripts bound to a bead are extended, primed by the unique barcode on the bead. Next, an endonuclease digests unused primer sites and the cDNA is PCR amplified. The sample is analyzed using a BioAnalyzer to inspect the size distribution and amplification extent of the cDNA pool. This pool is then prepared for sequencing using a ‘tagmentation’ process and samples are sequenced using paired-end sequencing on an Illumina sequencer. The first read reads off the special Drop-Seq barcode, and the second read provides the sequence of the 5’ end of the original transcript. This is known as 5’ tag counting, which is discussed in the Amplification Bias section(78).

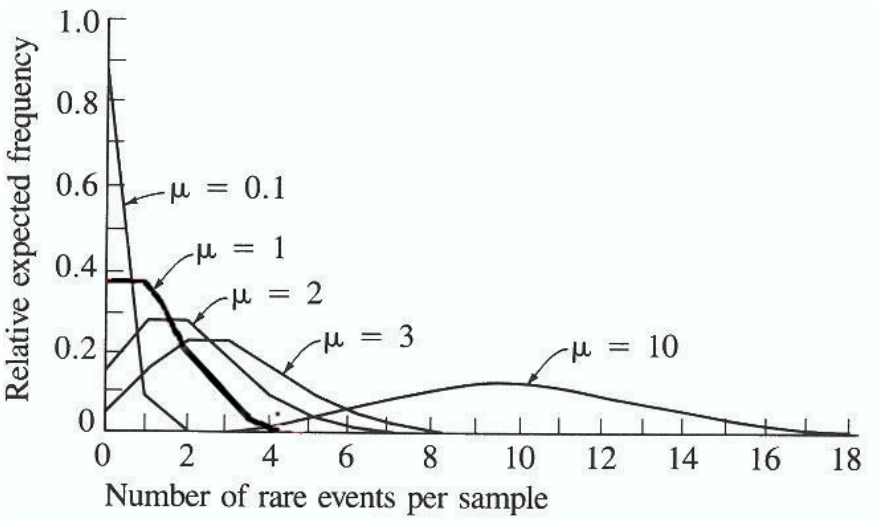


Figure 5: Diagram of Poisson probability distributions : Beads are encapsulated at very low values of μ ~ 0.1 such that on average less than 1/10 cells receives a bead. This ensures that the rate of doublets or more, i.e. events per sample > 1, where an event represents encapsulation, is very close to zero.

Sequencing results are then processed using software provided by the Drop-Seq authors to demultiplex, map, and count the transcripts. Briefly, the process involves extracting out the barcode portions of the read, trimming primer sequences, aligning the reads to a reference genome, selecting true STAMP barcodes, and creating a digital gene expression (DGE) matrix. The DGE generation process uses both the cell and the UMI components of the barcode. The cell barcode is used to demultiplex cells while the UMI to control for amplification bias during the multiple amplification steps.

As opposed to the C1 system where the C1 chip can be visually inspected for cells doubling up in a well, in Drop-Seq there is no way to physically inspect droplets as cells as lysed shortly after creation of the droplets. Thus, the protocol is validated by conducting a species mixing experiment. In the species mixing experiment, a 50:50 suspension of mouse 3T3 cells, and human HEK293T cells is flowed through the system. Sequencing results are demultiplexed and transcripts are counted for

each bead. As the counting tags each read with a cell barcode, cell barcodes are inspected to determine the percentage of transcripts that align to the mouse genome versus the human genome. Ideally, cell barcodes should be associated with only mouse transcripts, or only human transcripts. However, by chance, sometimes two cells will enter a droplet together and in the event that it is a human cell and a mouse cell, the system will be able to easily detect this occurrence. The frequency of these occurrences provides an estimate of how often two cells of any type are co-encapsulated, and thus how robust the system is at generating true single cell droplets. In a successful experiment given standard cell loading concentrations, fewer than 5% of STAMPs should contain both human and mouse transcripts(77).

Comparison of Single Cell Methods and shortcomings

The C1 and Drop-Seq provided powerful methods for reproducible, multiplexed RNA-sequencing analysis, yet each has its own strengths and weakness that must be considered.

The C1 is inherently lower throughput than Drop-Seq, as the commercially available chips at the time of experimentation could capture a maximum of 96 cells with single cell capture rates close to 70%, resulting in typical capture of approximately 70 cells(28). This becomes most significant from a cost and timing perspective as the cost for capture of a single cell on the C1 system is approximately $3.50 whereas for Drop-Seq it is $0.10 depending on the cell density and acceptable doublet rate(77). From a timing perspective, a typical Drop-Seq run captures between 200 to 2,000 STAMPs in a 15-minute interval. An hour of runtime can thus produce up to 8,000 STAMPs in the same amount of time as it takes to run the

capture step on the C1(28, 77, 79). On Drop-Seq, however significant personnel time and expertise is required to process the samples after collection of the droplet emulsion which can introduce significant technical variability as many of the steps need to be performed quickly and carefully. The C1, on the other hand, requires much less hands-on time and the protocol is less technically demanding.

The C1 has the distinct advantage of being able to visualize captured cells in the well before they are lysed. This allows for determination of wells containing true cell doublets, or empty wells. However, this visual identification was found to be faulty in several models of chips by Fluidigm after a species mixing experiment revealed significant doublet populations on visually confirmed single cells(77). Fluidigm determined this was a flaw in their well design that allowed two cells to sit on top of one another in a way that prevented discriminating individual cells(80).

While the species-mixing experiment confirms the overall viability of Drop-Seq for generating single cell libraries, it provides no guarantees for doublet formation in subsequent experiments. Notably, in experiments involving digestion of tissues, incomplete digestion of cell-cell proteins can result in significant doublet artifact with potentially minimal ability to detect these doublet populations. Visual identification of cells does not, however, eliminate the issue of ambient RNA contamination of results which is a significant problem for all single cell sequencing protocols, especially for sequencing of tissue samples as they all involve generation of a cell suspension.

“Amplification bias” refers to the technical variability introduced by PCR amplification of transcripts due to their size, GC-content, and potential secondary structure(81). In the cause of transcriptional analysis, this causes certain transcripts

to be more easily amplified during cycles of PCR relative to others, and because the process results in exponential amplification, two transcripts present in relatively equal abundance, may be amplified such that one is orders of magnitude greater abundance than the other.

Unique molecular identifier (UMI)-counting was first used by Kivoja et al to address general amplification bias(82), and then later applied directly to single-cell RNA-seq by Islam et al.(78) Amplification bias is especially significant in single cell samples where significant amplification cycles are required to bring the small quantities of RNA in a single cell into the detectable range. A UMI strategy for single-cell sequencing requires presence during RT of a randomly barcoded primer consisting of 5-8 bases resulting in 1,024 – 65,536 unique barcodes, depending on the length. These barcodes are inherited during first-strand synthesis of cDNA and are maintained during all cycles of amplification. During paired-end sequencing, the UMI barcodes of all the transcripts are read from one end, and the gene identity is read from the other end. It is assumed that all detections of a gene-UMI combination resulted from amplification of one original molecule of mRNA. Thus, detected events of a gene-UMI combination are collapsed to a count of one. In the case of Drop-Seq, a set of reads with identical cell barcode-gene-UMI reads are collapsed to an expression value of one for that gene in that cell barcode. Introduction of UMIs has been shown to significantly reduce the technical variability and increase reproducibility in single cell experiments relative to simply counting reads(78). The trade-off is that the 5’ end of the original mRNA molecule most distal to the polyA priming site is lost and thus only the 3’ end of a molecule is detected. This limits the

ability to detect mutations, allele specific transcripts, and alternative splicing that occurs at the 5' end of the mRNA.

Drop-Seq implements both cell barcoding and UMI counting, thus allowing for correction of amplification bias. The C1 protocol used in this paper does not use UMI counting, and thus allows for sequencing of the full transcript, but the data do suffer from noise introduced by amplification.

The newest platform for single cell RNA-sequencing is known as the Chromium 10x system. The system is akin to the Drop-Seq in that it uses barcoded beads to capture cells in nanoliter droplets. The advantages to this system are lack of plug-and-play operation and significantly improved cell capture. The system does not require assembly as the Drop-Seq rig does and the RT process happens inside of the droplet obviating a major ambient RNA contamination concern. The Chromium system also captures approximately 50% of the input cells because there is no possibility for bead doublets and thus the system only needs to operate at Poisson-limiting cell quantities. Despite these advantages, the device and consumable remain quite expensive such that the per cell isolation cost is still approximately an order of magnitude higher than Drop-Seq.

Statement of Purpose

The purpose of this line of research was to discover heterogeneity of well-studied models of IPF on a single-cell basis.

Specifically our goals were to:

1. Analyze the shift in cell populations and their expression profiles using multiple *in vivo* models of fibrosis on the C1
    a. Bleomycin mouse model
    b. Tgfβ1 induced fibrosis model
2. Use Drop-Seq to determine the heterogeneity of response to FENDRR knockdown on a single cell method, as well as to determine the usefulness of Drop-Seq for future experiments.

We hypothesized that in our mouse models, we would see a significant upregulation of apoptotic and fibrotic markers in bleomycin treated and Tgfβ1 induced mice in a single cell specific manner. Specifically, we expected to see single cells that demonstrated an expansion in the number of inflammatory cells, fibroblasts and myofibroblasts, and potential transitional phenotypes.

In the FENDRR knockdown experiment, we expected to see each cell demonstrate a clear discrimination of response to siRNA treatment. We also hypothesized there may be some population heterogeneity of both samples that we would be able to detect due to differences in the cell cycle state of the cells. Finally, given the importance of FENDRR as a regulator of cell cycle state, there may be distinct changes in the impact of FENDRR knockdown depending on the state of the cell during FENDRR treatment.

Overall these experiments provide internal validation for future use of these cutting-edge methods in larger mouse experiments and clinical samples.

Methods

1) Generation of model mouse systems

Doxycycline-inducible Tgfβ1 *transgenic mice.* The triple transgenic *CC10-rtTA-tTS*-Tgfβ1 (hereby referred to as Tgfβ1+) mice were previously generated by the Elias Lab(17) and key details are summarized here. Three transgenes were knocked-in to C57BL/6 mice using pronuclear injection and characterization of the progeny. The first construct contains the lung-specific CC10 promoter and a tetracycline inducible Tet-O operon activator (rtTA). The second contains the tet-O operator, a minimal CMV promoter, and a modified Tgfβ1 cDNA. Crucially, the Tgfβ1 is modified substituting serine codons for cysteine at positions 223 and 225 preventing latency associated protein-binding, and thus fully active Tgfβ1. The third construct contains a tetracycline controlled transcriptional silencer (tTS) driven by the CC10 promoter. The CC10 promoter drives lung-specific production of the rtTA and rTS, the former is active under doxycycline, and the latter is inactive under doxycycline. The presence of dox liberates the rTS from the Tet-O operon and allows binding of rtTA thus allowing for a tightly controlled and reversible lung-specific system for the expression of Tgfβ1.

In the experiment, two female 6-8 week old Tgfβ1+ mice, one with doxycycline added to the water, and the other without doxycycline added, and one wild-type littermate was used as a control. Doxycycline was added to drinking water at a concentration of 0.5 mg/mL and administered from day 0 to day 7. Mice were sacrificed using a lethal dose of urethane and tissues harvested at day 14.

Bleomycin-induced lung injury mouse model. One wild-type C57BL/6 mice at 6-8 weeks of age was given one dose of intratracheal bleomycin at an approximate dose of 0.5U/kg on day 0. One control mouse was treated with intratracheal PBS as a control. Mice were sacrificed using a lethal dose of urethane and mouse lungs were harvested at day 14 employing perfusion of the lung with PBS via the right ventricle.

2) Tissue extraction and dissociation

Extracted lungs are dissociated using the MACS automated dissociation system (Miltenyi) according to the manufacturer protocol. Briefly, extracted lungs were thoroughly washed in a PBS solution and placed in gentleMACS C tube containing a proprietary enzymatic digestion solution. A fast round of mechanical digestion was followed by slow agitation at 37C for 30 minutes, and another fast round of mechanical digestion. Cells were pelleted by centrifugation, re-suspended, and filtered twice to remove dead cells and debris.

3) C1 based library creation

C1 library creation was achieved according to the manufacturer manual. The medium size IFC chip was used as it is encompasses the expected size range of mouse lung cells (10-17μm). The IFC was primed by loading Harvest Reagent, Preloading Reagent, Blocking Reagent, and Cell Wash Buffer into specified wells on the IFC chip and placed into the C1 system for 12 minutes to complete the priming protocol. Cells obtained from tissue extraction were re-suspended in Suspension Reagent to a final concentration of 100/μL and 60μL was loaded onto the inlet port of the IFC. The IFC was placed into the C1 system for 65 minutes to run the loading protocol. After loading, the wells were individually inspected to determine the

occupancy (0, 1, 2 or debris). In the next step, Harvest Reagent, lysis mix, RT mix containing SMARTScribe Reverse Transcriptase (Clontech), and PCR mix containing the Advantage 2 Polymerase Mix (Clontech) were loaded into the designated inlet ports, and the IFC was placed into the C1 for the 8.5 hour process of lysis, reverse transcription, and harvest. Notably, the SMART-seq2(83) RT strategy was used and 21 cycles of amplification were used as per the standard C1 protocol. Amplified products were transferred to a 96 well plate and a modified Nextera XT (Illumina) DNA library preparation protocol was carried out. Post-amplification cDNA were quantified individually using a multiplexed Picogreen assay and read on a plate reader to determine the required dilution for tagmentation. Samples were diluted to 0.2ng/μL and incubated with tagmentation buffer that simultaneously fragments the full length cDNA and tags the fragmented molecules with a standard Illumina barcode. These fragmented molecules were then incubated with Illumina specific multiplexing index primers and another 12 cycles of PCR were carried out. The samples were pooled and AMPure XP beads (BD) used at a ratio of 0.9:1 of pooled volume to remove primers. The final pooled library was quantified using a BioAnalyzer High Sensitivity DNA chip (Agilent). Sequencing was performed on a HiSeq 2500 (Illumina) in High Output mode.

4) NHLF FENDRR knockdown (KO)

siRNA targeting duplexes were designed using the web tools available from GE Dharmacon. The sense sequence was ‘GAAGAUACCAAGUGAAAUAUU‘ and the anti-sense sequence was ‘UAUUUCACUUGGUAUCUUCUU’. Manufacturer designed and validated non-targeting siRNA were used as a negative control.

Transfection of siRNA was conducted with Lipofectamine 2000 diluted in Optimem (Life Technologies) reduced media. After 24h cells were harvested from the plate and prepared for the Drop-Seq protocol.

5) Drop-Seq based library creation

Drop-Seq protocols were executed using the reagents, primers, and protocols described in the Drop-Seq 3.1 manual and paper(77). An overview of the steps of the protocol is provided in the introduction as an overview of the Drop-Seq protocol and further details about each experiment are provided below.

Species Mixing Experiment

Mouse 3T3 cells and human HEK 293T cells were cultured using standard conditions to 75% confluence and harvested by TrypLE. Cells were diluted to a final concentration of 100 cells/μL in and barcoded beads (Chemgenes) were suspended in lysis solution to a final concentration of 120 beads/μL. Cells, beads, and oil were loaded onto syringe pumps and flowed at a rate of 4,000, 4,000, and 15,000 μL/hr respectively to create a droplet in oil suspension. After droplet generation, droplets were broken using a 30mL solution of perflurooctanol in 6X SSC. The recovered beads, a subset of which were heavily enriched with the RNA of the cell they were encapsulated with, were washed several times with 6X SSC to remove ambient RNA and resuspended in a RT mix containing Maxima H- reverse transcriptase (Thermo Fisher). Beads were washed again and incubated with Exonuclease I (NEB) to remove unextended primers on the beads. The beads were washed again, counted on a hemocytometer, and diluted to aliquot 2,000 beads into each PCR tube. A total of 6,000 beads were aliquoted for 10 cycles of PCR with HiFI Hotstart Readymix

(KAPA Biosystems). Note that the recommended cycle number for the species mixing experiment is 13 cycles. PCR products were pooled and cleaned up using AmpureXP beads at 0.6:1 ratio and the post-clean up products were characterized on a BioAnalyzer HS Chip. The pooled samples were tagmented and enrichment PCR using special 'P5-SMART-PCR hybrid oligo enables amplification of only tagmented fragments that contain the Drop-Seq barcode. 12 cycles of PCR were performed, and the sample is cleaned up using the 0.6x AmpureXP bead ratio and characterized on the BioAnalyzer HS chip. A final 10uL library pool at a 3nM concentration was created as the input for denaturation for running on the MiSeq platform. Read specification for the MiSeq were as follows: Read 1, Read 2, and Read 1 index lengths are 20bp, 50bp, and 8bp respectively.

FENDRR KO Experiment

The FENDRR KO experiment was conducted identical to the species mixing experiment with the following notable changes: Cells were loaded at a concentration of 50 cells/μL and beads were loaded at 120/μL. Sequencing was performed on the HiSeq 2500 in High Output mode.

6) Bioinformatics Analysis

Analysis pipeline for C1

Raw reads were mapped using the Kallisto(84) pseudoalignment software to generate transcript counts. Transcripts mapping to ribosomal proteins were removed by selecting out transcript names beginning with Rpl and Rps. Subsequent analysis was performed with Bioconducter package 'scater'(85). Multimapped transcripts mapping to the same gene were removed by keeping only the first instance of the

mapped transcript to avoid overrepresentation of specific genes. Only wells observed to contain single cells after initial the loading protocol were carried forward for further analysis. Low quality cells were then filtered out in order to eliminate cells with (1) total library size three median absolute deviations (MADs) below the median log-library size, cells in which the (2) log-transformed number of expressed genes is three MADs below the median, and (3) the proportion of transcript counts coming from mitochondrial genes greater than 10%. Subsequently, only features expressed in greater than 5 cells in each individual experiment were kept to eliminate noise generated from either imprecise mapping of low quality transcripts or low level stochastic expression. The union of the features remaining across each of the experimental groups was then combined, thus keeping only transcripts present in at least 5 cells in any of the experiments. Transcripts were then summarized at the gene level using the 'summarizeExprsAcrossFeatures' command which combines features using the observed transcripts per million. The expression levels were then normalized by using size factors computed by the 'computeSumFactors' command and setting the sizes of the pooling groups to ¼, ½, ¾ and 1x the total number of cells passing filtering. The normalized expression values were then fitted using LOESS regression and with Span=0.1 to estimate the biological and technical components of the variance of gene expression in each cell in an experiment. For unsupervised clustering, we selected a set of highly variable genes defined by those with biological variability significantly greater than zero at a specified FDR and biological variability threshold. A correlation matrix amongst these highly variable genes was constructed and an FDR threshold of 0.05 is used to select pairs of

correlated genes. The top 50 genes with the highest biological variability of this set of correlated genes was used to construct a dissimilarity matrix and hierarchical clustering was applied to create a heatmap. Potential clusters are identified using the package ‘dynamicTreeCut’(86) with a minimum cluster size of 10. For semi-supervised clustering, an internally derived list of marker genes was developed. And hierarchical clustering was used to group cells according to this marker gene list with a minimum cluster size of 10.

Analysis pipeline for Drop-Seq

Demultiplexing was performed using the Drop-Seq tools software(87) following the standard parameters in the Drop-Seq Informatics Cookbook v1.2(88). Briefly, the following steps were taken to prepare the raw reads into a ‘cell x transcript’ count matrix. FASTQ files were converted into query-name sorted BAM files using Picard(89). Then cell barcodes were extracted as the first 12 base pairs of Read 1 and transferred into metadata; a similar process extracted the next 8 UMI-encoding base pairs. Low quality reads were removed at this stage and primer sequences and polyA tails contaminating the reads were clipped away. The BAM file was converted back to a FASTQ file for alignment using STAR 2.5(90). The human and mouse genomes used in this analysis were the hg19 and mm10 reference genomes(91), respectively and the Ensembl release 87(92) genomic features files.

The alignments were merged with unaligned metadata tagged BAM files to tag all the aligned reads with cell and UMI barcodes. Aligned reads were also tagged with a gene name metadata tag if the read occurred in the region of an annotated gene. Next, a bead synthesis quality detection program was run on the cell barcode and

UMI to detect any potential barcode or UMI redundancy issues associated with improper pool-and-split bead synthesis.

The ‘knee-plot’ for cumulative reads was generated by plotting the cumulative sum of reads associated with each barcode, by descending order of number of reads per barcode. A cut-off was selected by inspecting the knee-plot for an inflection point were the number of reads per STAMP decreases significantly representing the zone of reads attributed to ambient RNA binding to beads. In the FENDRR KO experiment, 300 and 200 cells were used as the cut off for the S1 and F1 groups. For the species mixing experiment, 200 cell cutoff was used. These cells were designated STAMPs and were taken forward for creation of the DGE matrix. The DGE matrix data for the FENDRR experiment was further processed using the Bioconductor package ‘Seurat’(93). Cells with greater than 5% mitochondrial reads, fewer than 2000 detected genes, and UMIs greater than 10,000 were filtered out. For the T-SNE plots, we selected the top PCs (1-10) and a resolution parameter of 0.6 to identify clusters. The ‘FindMarkers’ command was used to identify the marker genes for each cluster; the top 10 most highly expressed of this group of these marker genes was used to construct a heatmap.

Results

1) Results of whole mouse lung experiments using the C1

a. Quality control results of C1 analysis

The overall capture rate for single cells for these experiments ranged from 44/96 to 90/96 with an average of 72 single cells captured (Table 1). In order to improve the specificity of detection of important cell groups, low quality cells and features expressed at low levels in only small numbers of cells were filtered out. This process improves the ability of clustering algorithms to robustly group cells and not be swayed by noisy, lowly expressed genes or outlier cells with low quality. Filtering was first performed to eliminate cells in which (1) total library size three median absolute deviations (MADs) below the median log-library size, (2) the log-transformed number of expressed genes is three MADs below the median, and (3) the proportion of transcript counts coming from mitochondrial genes greater than 10%. This process eliminates cells that deviate from the remainder of the pool in

Mouse Lung Sample	Total Reads	Single Cell Wells	Library Size Filter	Gene Diversity Filter	Mito Genes Filter	Final Cells
TGF-B Transgenic	145,190,871	90	1	4	26	**64**
Doxycycline control	141,510,510	42	0	0	16	**26**
Bleomycin treated	132,580,429	90	4	3	14	**71**
PBS treated control	164,771,402	78	3	2	24	**53**
Wild-type control	152,499,912	64	9	8	11	**46**

Table 1: Summary of read depth, single cell capture, and filtering steps. Results of quality filtering of whole lung samples on the C1 platform. Filtering criteria are described in depth in the text. Note that certain cells meet criteria for filtering and thus the sum of the filtered cells does not equal the total number of cells filtered.

terms of number of transcripts detected, diversity of transcripts detected, and the health of the cell based on the dominance of mitochondrial genes detected, respectively. In Figure 6, this process is shown in further detail. Under these filtering criteria, the final cell counts were extremely low for the doxycycline control at 26, and best for best for the Bleomycin treated mice.

Additional filtering was subsequently performed to remove lowly expressed transcripts that represent likely noise in the dataset. In our analysis, we removed transcripts that were not expressed in at least 5 cells. Transcript counts were

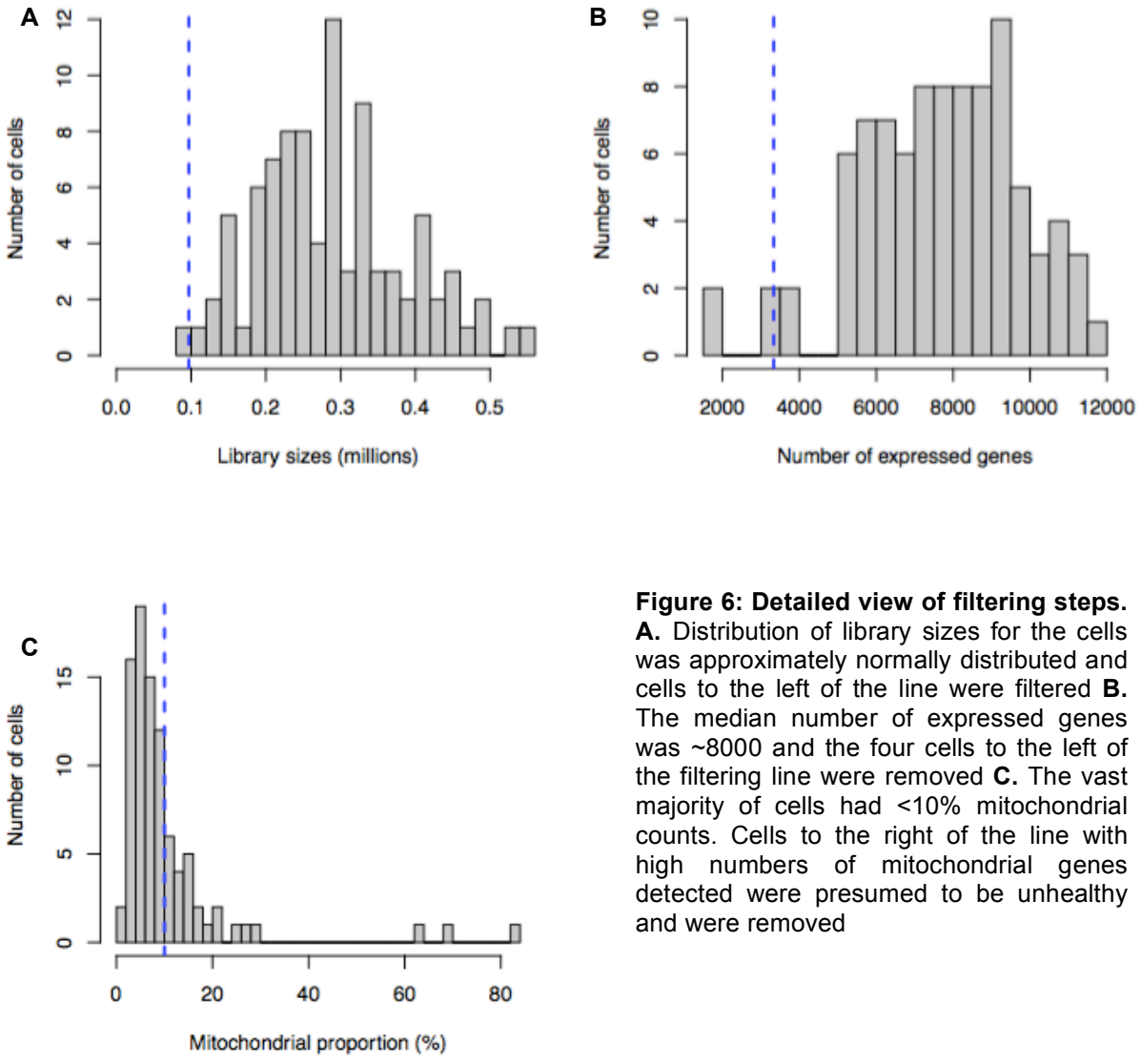


Figure 6: Detailed view of filtering steps. A. Distribution of library sizes for the cells was approximately normally distributed and cells to the left of the line were filtered **B.** The median number of expressed genes was ~8000 and the four cells to the left of the filtering line were removed **C.** The vast majority of cells had <10% mitochondrial counts. Cells to the right of the line with high numbers of mitochondrial genes detected were presumed to be unhealthy and were removed

aggregated at the gene level by summing over the transcripts per million of each transcript and then gene counts per cell were normalized. The post filtering distribution of the library size and the transcripts present (marked as 'features') in each cell are shown in Figure 7. The Tgfβ1 exposed mouse is distinct in that the cells all demonstrate a high diversity of transcripts and a tight distribution of the total transcripts detected.

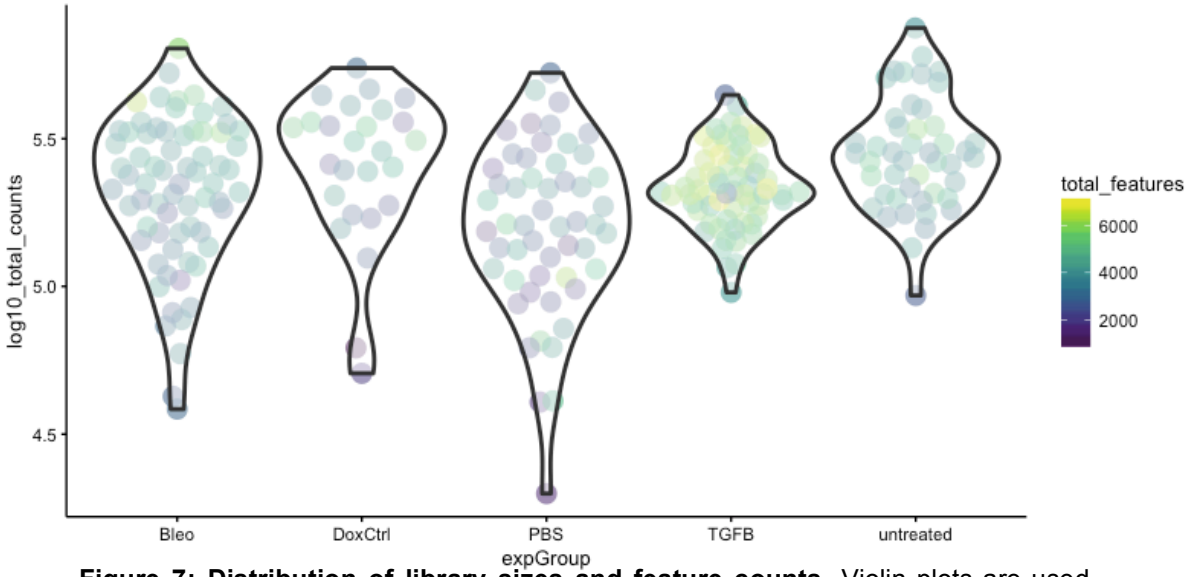


Figure 7: Distribution of library sizes and feature counts. Violin plots are used here to show the distribution of the library sizes of each of the experimental groups. Each circle represents a cell, its position on the y-axis represents the log-normalized number of transcripts detected for that cell, and the color scale represents the number of unique transcripts detected.

b. Assigning Cell Identity based on marker genes.

A set of marker genes was developed internally for canonical genes used to categorize specific cell populations in the lung. Specifically, this list encompasses Type I Epithelial (AT1), Type 2 Epithelial (AT2), general alveolar epithelial, endothelial, pericyte, fibroblast, monocyte, and general mesenchymal stem cell lineage cells. These marker genes were then used to cluster each experimental

treatment individually. Figure 8 shows the results of semi-supervised clustering for the Tgfβ1 data set based on this list of marker genes. The likely epithelial and mesenchymal lineages are indicated by the red horizontal bar and the combination of the purple, green, and cyan bars, respectively. There are several notable features of this data set. Under the purple cluster of cells are putative fibroblasts with higher expression of Col1a1 transcripts. A subset of these cells also expresses Acta2 and thus represents the myofibroblast subpopulation that is known to expand under fibrotic conditions. There is also a set of cells that expresses S100A4, which in humans is known as Fibroblast Specific Protein 1 (FSP-1) but does not express Col1a1 or Acta2. Finally, in the red cluster marked with high epithelial cell specific markers are a group of cells expressing Col1a1 and Acta2 at higher levels than the

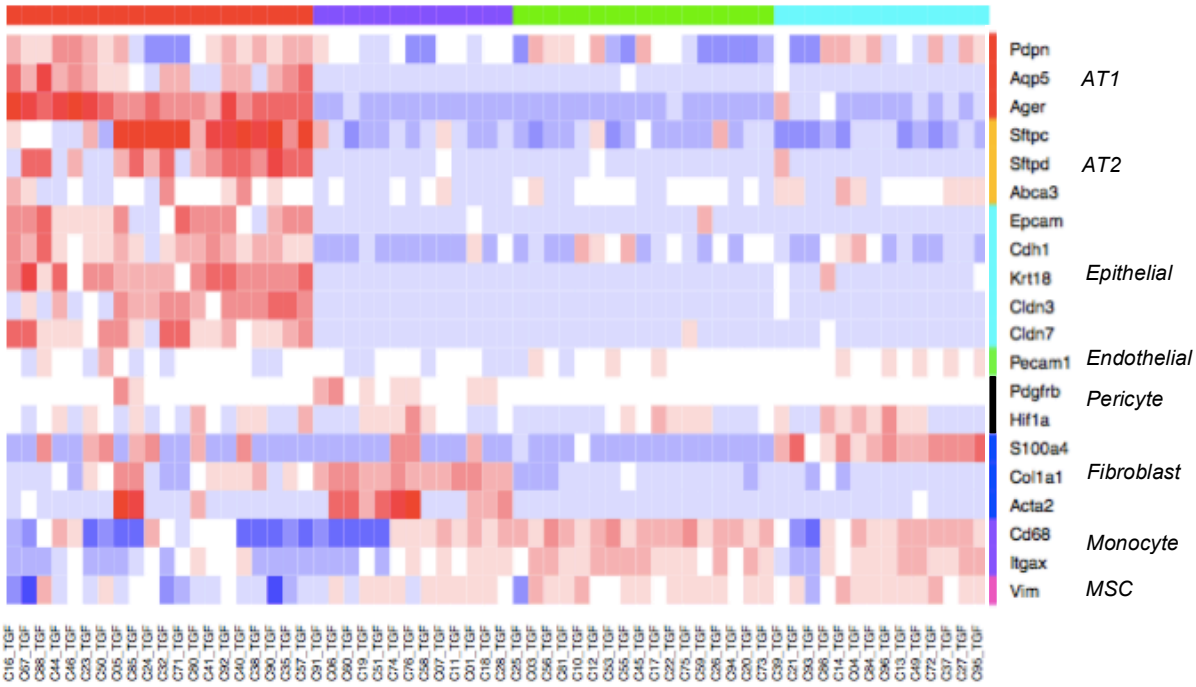


Figure 8: Semi-supervised clustering of Tgfβ1 **experiment based on marker genes.** Clustering for Tgfβ1 mouse lung cells shows clear populations of cells expressing epithelial markers (red horizontal bar) and fibroblast markers (purple and cyan bars).

remainder of the cells.

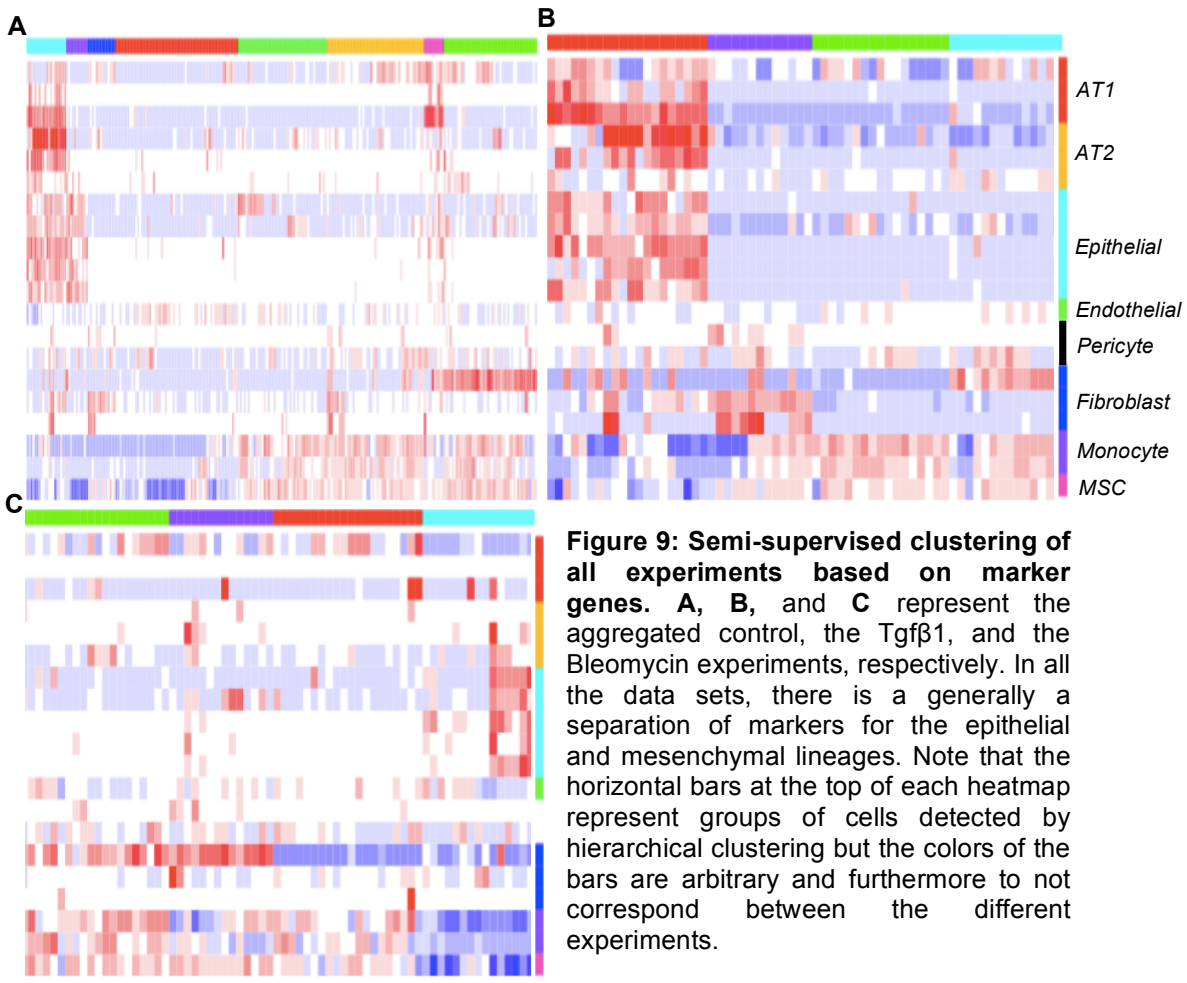


Figure 9: Semi-supervised clustering of all experiments based on marker genes. A, B, and **C** represent the aggregated control, the Tgfβ1, and the Bleomycin experiments, respectively. In all the data sets, there is a generally a separation of markers for the epithelial and mesenchymal lineages. Note that the horizontal bars at the top of each heatmap represent groups of cells detected by hierarchical clustering but the colors of the bars are arbitrary and furthermore to not correspond between the different experiments.

In figure 9, semi-supervised clustering for the Tgfβ1, Bleomycin, and combined control cells is shown. Generally the clustering bars along the top of the pooled control and bleomycin-treated mice also show separation along the epithelial-mesenchymal axis. As expected, the number of cells showing fibroblast markers is most expanded in the bleomycin group in this data set, and is small and not strongly expressing fibroblast markers in the control data set.

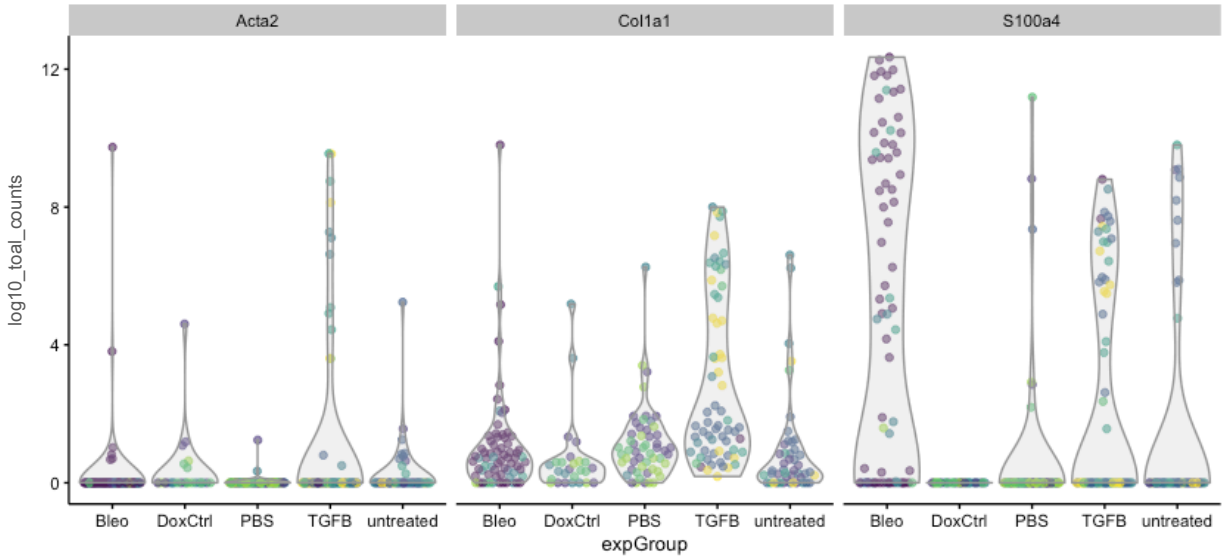


Figure 10: Expression of fibroblast specific markers. Violin plots are used here to show in more detail the log-normalized differences in Acta2, Col1a1, and S100a4 between the treatment groups. Each point represents a cell, and each label on the x-axis denotes the sample that the cells shown in the plot above the label came from. Cells are colored based on the cluster they fall into using our unbiased clustering approach. Note, that the clustering colors are consistent within each treatment group but have no meaning in between different groups.

c. Unbiased clustering of cell types

Cells were also analyzed using an unbiased clustering methodology. Similar to the semi-supervised clustering results, the Tgfβ1 experiment demonstrated the most robust clustering. Using a minimum cluster size of 10 cells, 6 clusters are identified with 4-5 gene modules. In figure 11, the top 50 most variable genes used in clustering are shown. The distribution of cells here is not representative of the normal lung environment(30). However, given the significant epithelial apoptosis and inflammatory infiltrate present in the characterization of this mouse(17), the distribution of cells is likely representative of the tissue environment in this mouse.

The same process performed on the aggregate control group and the bleomycin group does not cluster in a robust fashion nor do the clusters represent sets of genes with biologically discernable correlated expression.

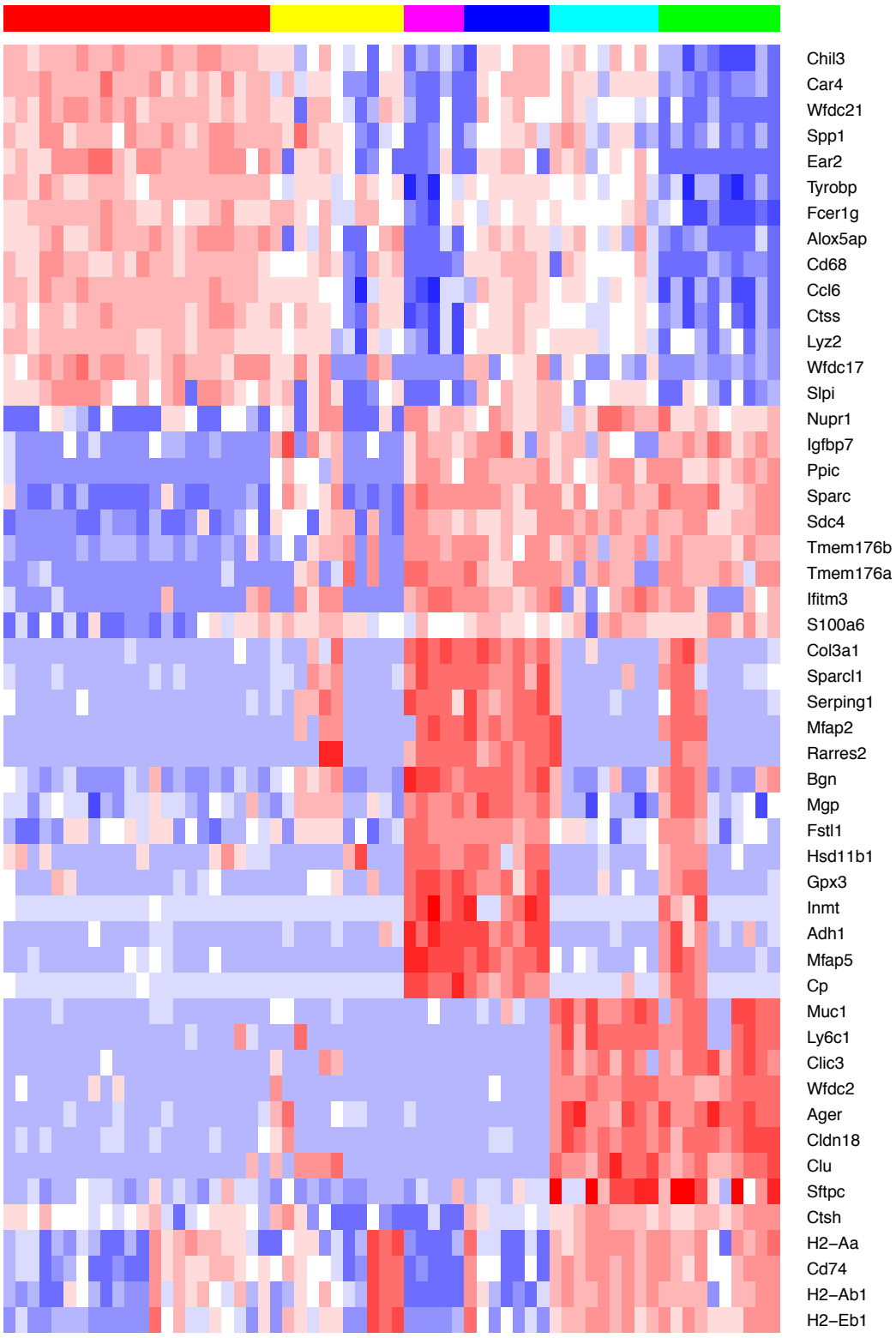


Figure 11: Unsupervised clustering of Tgfβ1 **experiment.** Hierarchical clustering of the dissimilarity matrix between cells is presented here. Using a minimum cluster size of 10, six distinct clusters are identified. Inspection of the heatmap shows four to five sets of genes on which these cells cluster.

2) Drop-seq analysis

a. Species mixing experiment confirms viability of platform to generate valid single cell libraries

The species mixing experiment was carried as described in the methods section. Initial experiments demonstrated excessive amplification of the post-PCR library (Figure 12). Subsequent libraries were prepared using the same set of beads and amplified using three fewer cycles. This resulted in properly amplified libraries similar to those in the DropSeq lab manual in Figure 12.B(88).

In Figure 13, we demonstrate successful implementation of the Drop-Seq platform to create single cell libraries. In Libraries 1 and 2, 1.9% and 2% of the cells

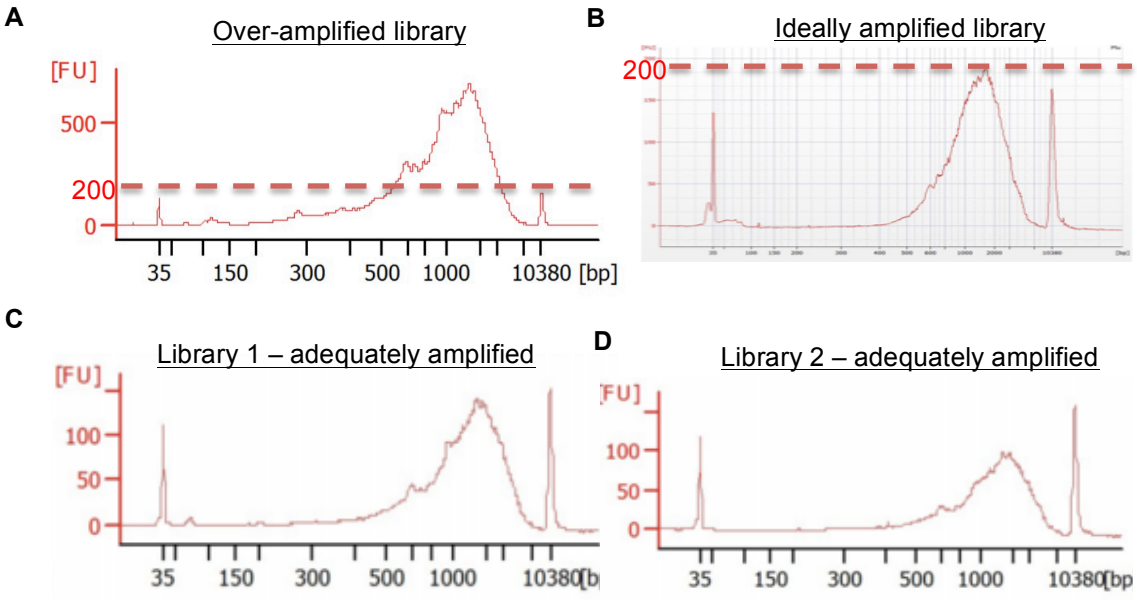


Figure 12: Amplification cycle determination for human-mouse experiment. Bioanalyzer provides the distribution of DNA size fragments found in a library. The x-axis represents the size of the fragment and the y-axis is a fluorescence measurement indicating the absolute abundance of fragments of the size indicated on the x-axis. **A.** An overamplified library we generated using the 13 cycles recommended in the Drop-Seq manual, this can be seen in comparison to **B.** which is reproduced from the Drop-Seq lab manual and is an ideal BioAnalyzer trace with an amplification peak near 200 FU and a peak fragment size near 1100bp representing the average size the reverse transcribed cDNA. We adjusted the cycle number to 10 and generated properly amplified libraries in **C. D.** with a peak near 100 FU at approximately 1100bp.

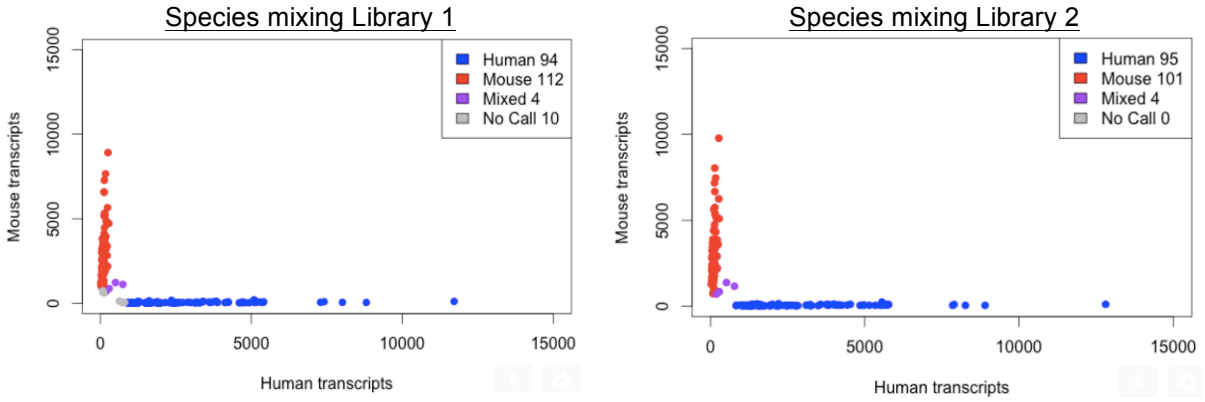


Figure 13: Results of species mixing experiment. Replicates of the species mixing experiment demonstrating successful implementation of the Drop-Seq platform. Each dot represents a STAMP and are shaded blue or red if they contain greater than 95% of transcripts from human or mouse, respectively. A purple cell represents barcodes with both human and mouse cells.

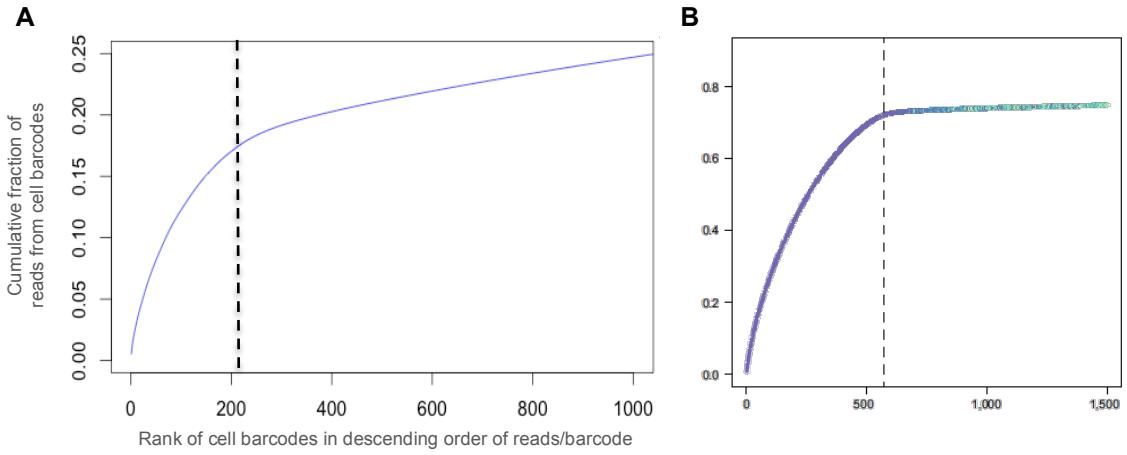


Figure 15: Inflection plots of species mixing experiment. X-axis is the rank of the barcode (in descending order) by the number of reads that barcode has. Thus the barcode represented by '1' has the most number of reads, '2' has the second most, and so on. The number of reads associated with each barcode flattens out and this flattening represents the region of ambient barcodes**. A.** Library 1 demonstrates poor segregation of reads between STAMPs and barcodes bound to ambient RNA. **B.** Right plot is from the original Drop-Seq paper. A sharp inflection point is seen demonstrating a clear boundary between bona-fide single cell reads and reads from ambient RNA. Note the difference in the scales.

contained a mixed population of transcripts indicating a doublet. The estimated doublet rate in the system at this cell concentration of 100 cells/μL is approximately 4% if the probability of different doublet combinations is accounted for. Indeed some outlier cells with greater than 10,000 transcripts are seen in both libraries. It is highly likely that these are human-human doublets.

The inflection plots for the species mixing experiment (Figure 15) demonstrate low capture of STAMPs relative to ambient RNA. The Drop-Seq authors demonstrate 70% of their reads originate from STAMPs versus 17% for Library 1, and 15% for Library 2 (data not shown). Moreover, there is a sharp inflection point in the Drop-Seq author data (77) as compared to the soft inflection point seen in our results resulting from ambient RNA contamination causing attachment of RNA to beads that were not encapsulated with a cell after droplet breakage.

No further analysis is performed on the species mixing experiment as the goal of this experiment is to provide system validation.

b. FENDRR KO Library is a high quality library

The inflection plots for the FENDRR experiment (Figure 16) demonstrate there was significantly greater capture of true cells as a fraction of the total number of reads. This indicates less ambient RNA due to a higher quality cell suspension or more clean library preparation In these libraries, the inflection point occurs near at

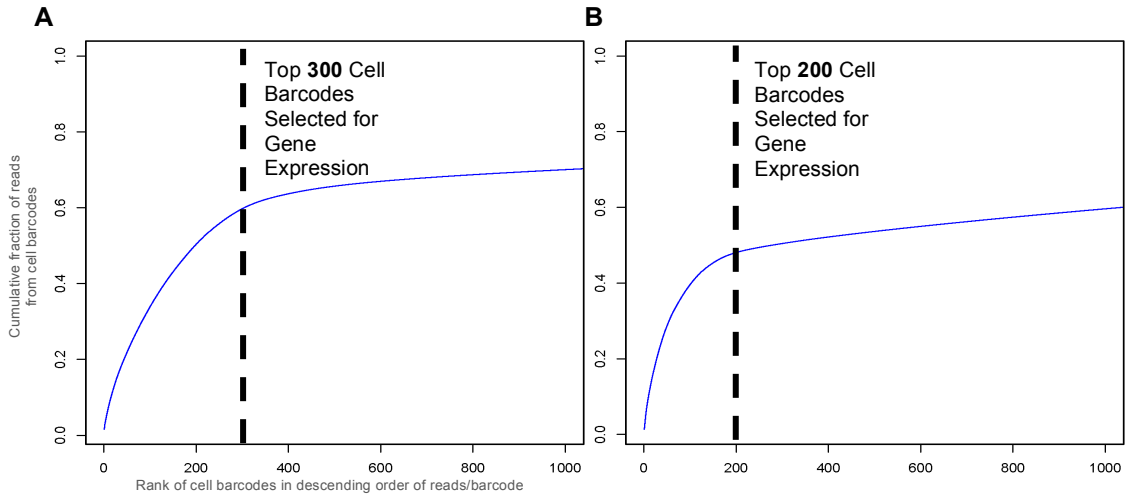


Figure 16: Inflection plots for FENDRR experiments. A. Inflection plot for the FENDRR incubated with scrambled, non-targeting RNA. **B.** FENDRR incubated with targeting siRNA inflection plot shows the sharpest inflection indicating the cleanest library preparation for this tissue and potentially as a result of greater resistance to apoptosis due to treatment.

200 cells for the control cells, and 300 cells for the FKO cells. After first pass filtering for low quality cells, not meeting a minimum expressed gene count of 2000 our data set consisted of 294, and 158 cells for the control and FKO groups. UMI counts are represented as a fraction of the total number of UMIs in the cell and multiplied by 10,000 and log transformed.

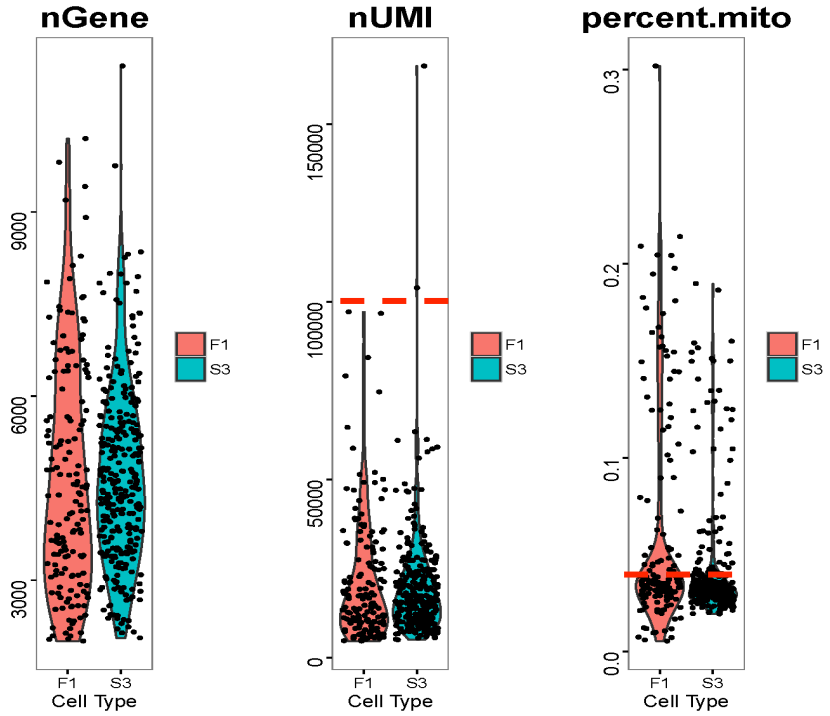


Figure 17: Quality filtering for NHLF FKO experiment. Read distributions of genes, UMIs, and mitochondrial reads. F1 represents the FENDRR KO cell and S3 represents the scrambled library.

Further quality filtering steps were applied to this data set (Figure 17). In the total population of NHLFs the number of genes detected averages around 5,000 with the minimum number of genes expressed at 2,000 as set by the filter, and tapered off near 9,000. 10,000 mRNA bindings is an apparent maximum of the binding kinetics when one cell is encapsulated with one bead. A second cell entering a droplet shifts the kinetics towards greater binding of mRNAs. Thus STAMPs with over 10,000 UMIs were also filtered. STAMPs with greater than 5% mitochondrial reads were also filtered as these indicate cell releasing mitochondrial RNA indicative

of increased cellular stress. This is a more stringent threshold than the total lung experiment because the UMI counting enabled by the Drop-Seq experiment adjusts for the potential amplification bias of mitochondrial genes.

c. FENDRR KO experiment confirms fibroblast changes seen in whole tissue experiments

In order to confirm FENDRR knockdown in the FKO population, we looked specifically at FENDRR as well as ACTA2, the gene coding for SMA. In our lab, we validated that FENDRR knockdown increases the abundance of SMA (unpublished data) and thus focused on this change in analyzing the Drop-Seq results. In Figure 18, FENDRR expression follows a bimodal distribution in both populations with a large number of samples showing no FENDRR expression. In the control population, this is likely due to dropout as FENDRR is expressed in relatively low numbers given

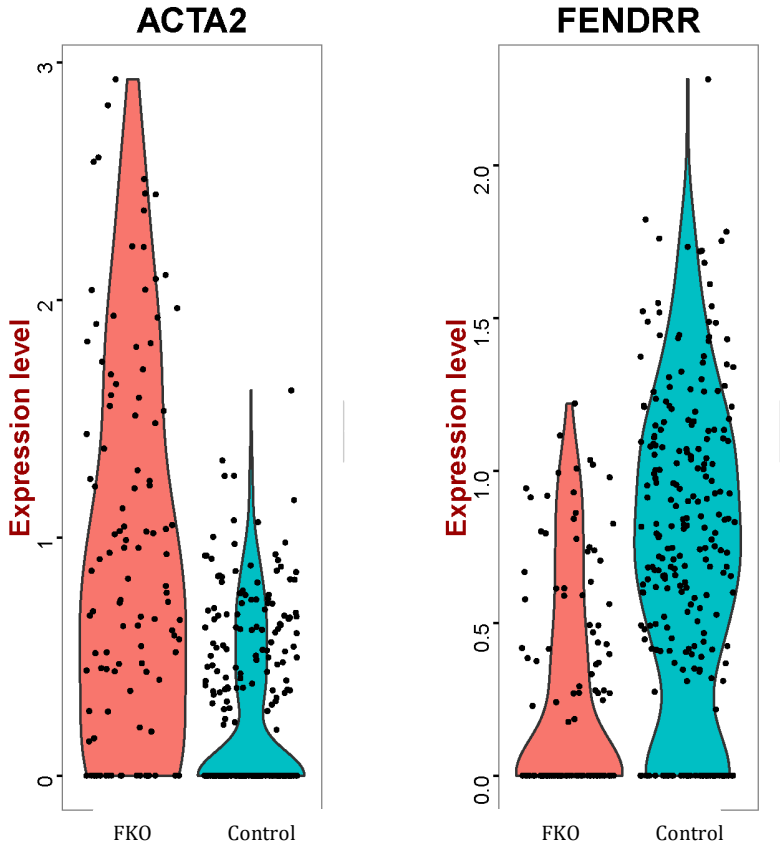


Figure 18: ACTA2 and FENDRR expression in FKO and control cells.

bead saturation kinetics, Drop-Seq has lower sensitivity for genes expressed at low levels. On average, FENDRR expression is decreased in the KO population. The cells still expressing FENDRR were likely due to inefficient transduction of FENDRR targeting-siRNA. ACTA2 is upregulated in the FKO population confirming the expected effect of treatment. There remains a population of cells that either do not express ACTA2, or express it levels similar to the control population.

PCA analysis was performed on the combined group of filtered cells and genes to understand the spectrum and magnitude of changes induced by FENDRR KO (Figure 19). A first set of plots was generated using all the genes resulting in a moderate separation of the two groups of cells and a wide distribution of cells in the control group. A set of highly variable genes was selected in order to remove noisy

Figure 19: PCA plots of FENDRR KO experiment cells. **A.** PC1 vs PC2 for all genes. **B.** PCA analysis using only the top 100 for (left to right) PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3.

genes that skew creation of principal components. PCA plots generated for the first three PCs demonstrate improved clustering of the groups in terms of their separation and their intergroup variability although the control cells remain more widely distributed than the FKO cells.

t-distributed stochastic neighbor embedding (T-SNE) based clustering was performed next as this dimensionality reduction technique is capable of accounting for non-linear relationships in the data(94, 95). We applied a T-SNE based clustering algorithm to this data set and three clusters were assigned. As seen in Figure 20, 'Cluster 3' is almost completely composed of the FKO cells. Figure 21 shows the side by side expression of ACTA2 and FENDRR in on the T-SNE dimensions and shows the inverse relationship between the expression levels of these two genes further confirming that high expressing FENDRR cells correspond to lower levels of ACTA2 expression.

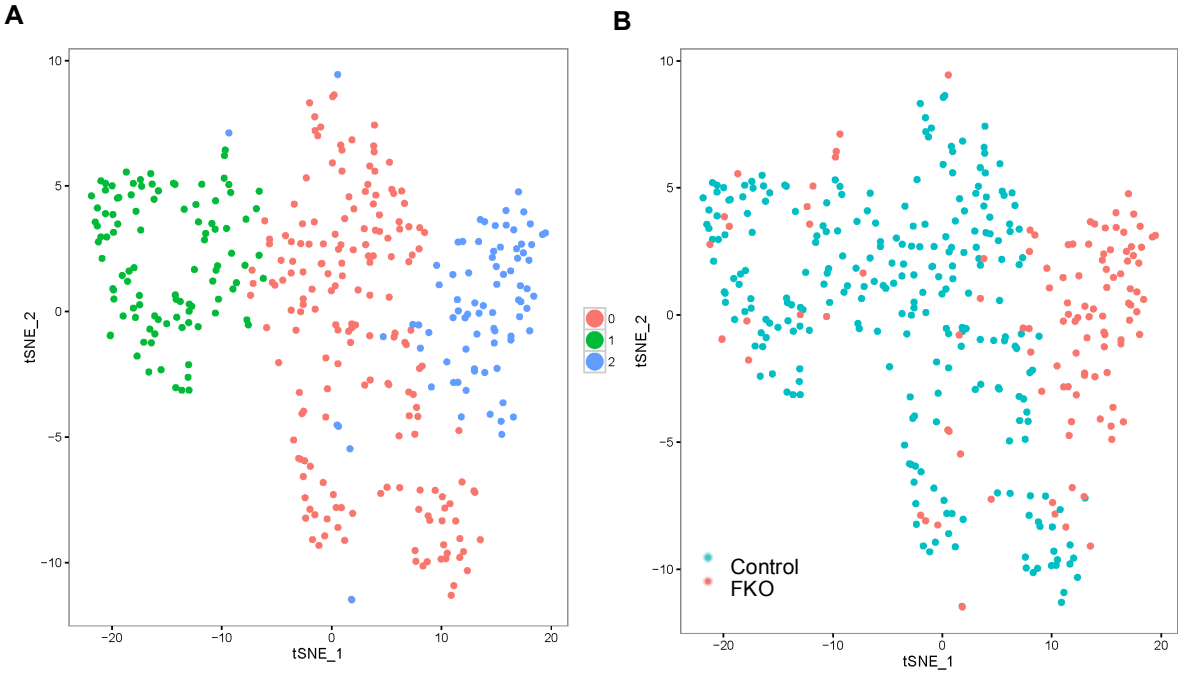


Figure 20: **T-SNE plot of FENDRR KO experiment. A.** The T-SNE plot clustering algorithm determines three distinct clusters. **B.** Shows the same configuration of cells as clustered by the T-SNE algorithm, but colored by the identity of the experiment.

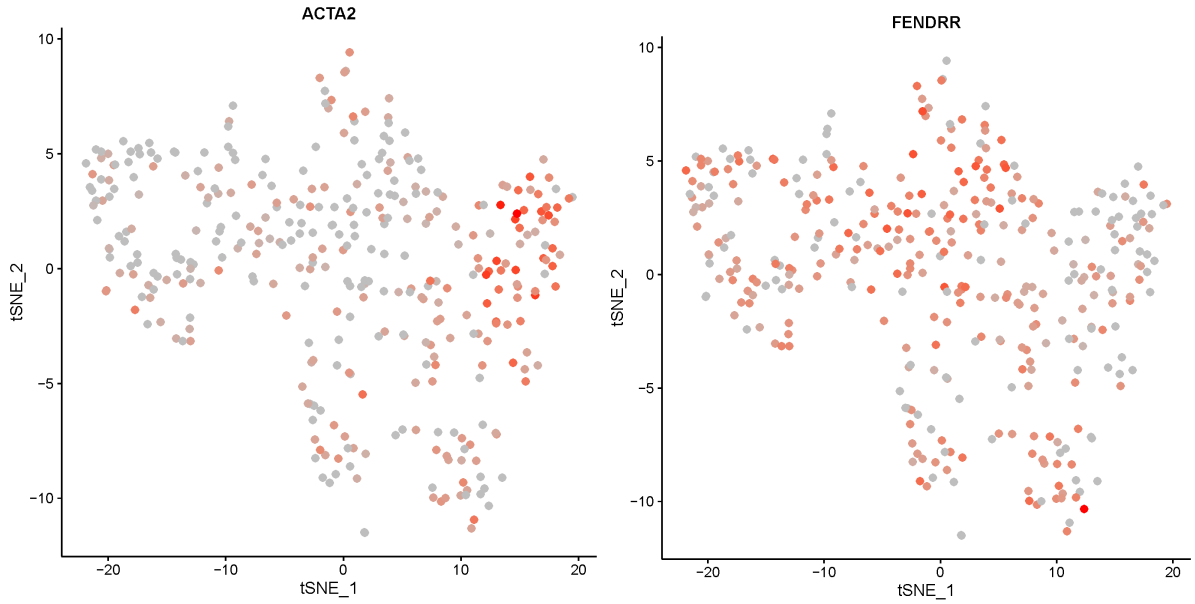


Figure 21: ACTA2 and FENDRR expression in the T-SNE generated clusters. The inverse relationship between low FENDRR expression (low expression in gray) and high ACTA2 expression (high expression in red), and vice versa, is more clearly shown on a cell specific basis.

In order to understanding the major drivers of the three clusters including the differences between ‘Cluster 0’ and ‘Cluster 1’ we selected the top 10 highly expressed genes in each cluster and created a heatmap (Figure 22). The top genes expressed in Cluster 1 are highly associated with G1 phase events (FDR = .0017) according to Reactome analysis(96, 97). Cluster 0 genes are not significantly associated with any process but are correlated with extracellular matrix formation and thus may correspond to features of normal fibroblast activity. Genes upregulated in ‘Cluster 3’, the cluster representing the FKO cells, are highly associated with a fibrotic phenotype including smooth muscle actin, a myofibroblastic protein(98), CDKN1A encoding p21 which has been found to be necessary for fibroblast entry into quiescence(99), and GDF15 has been show to regulate fibroblast

activation(100) leading to increasing expression of smooth muscle actin in cancer associated fibroblasts and paradoxically in this scenario, proliferation(101). IGFBP5 also induces a fibrotic response associated with EMT and Tgfβ1 upregulation(102).

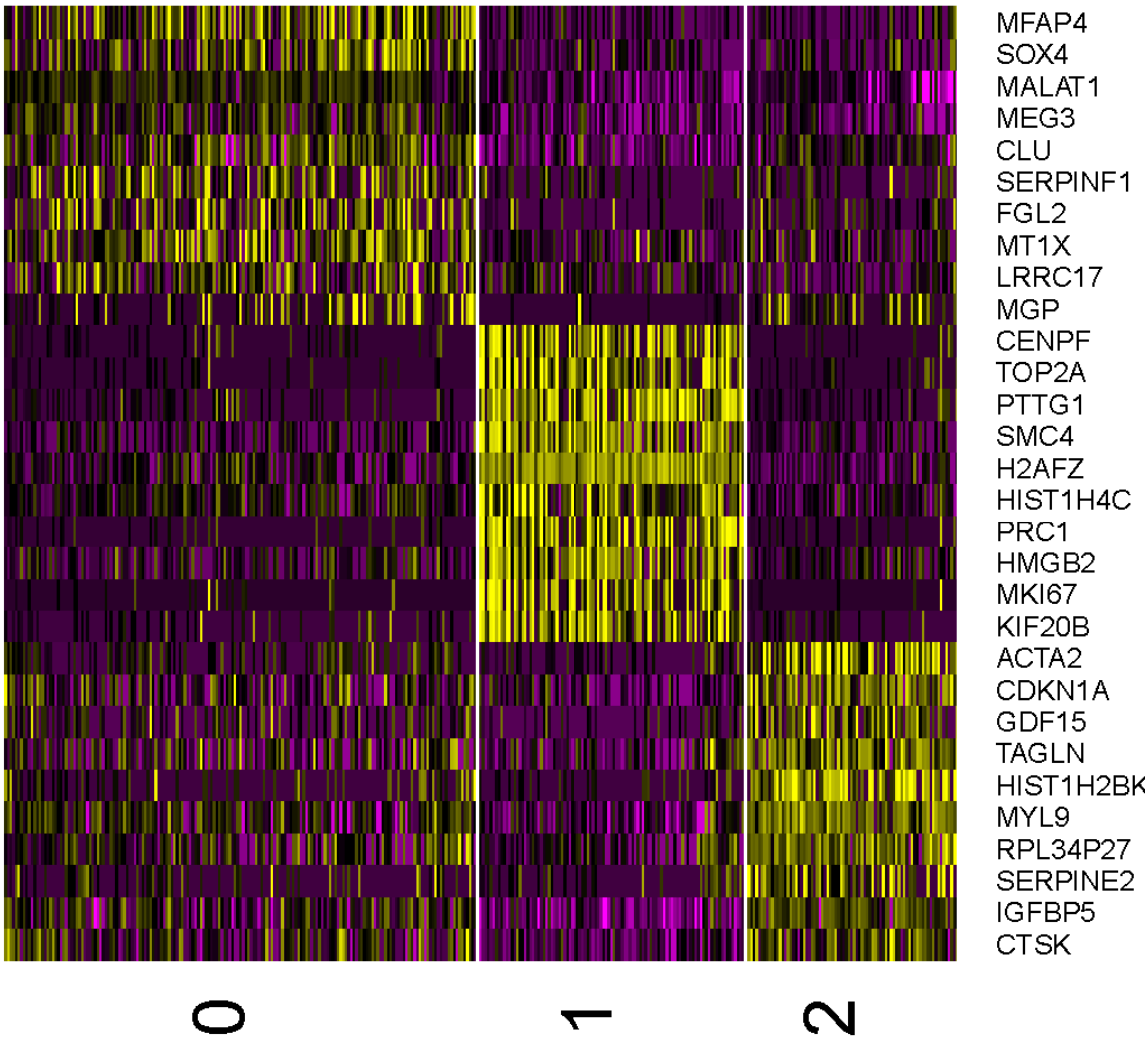


Figure 22: **Heatmap divided by clusters generated by T-SNE analysis.** The top differentially expressed genes in each cluster are shown in the heatmap revealing expected relationships due to FENDRR treatment and cell cycle status of the cells.

Discussion

We present here results from two single cell experiments. In one experiment, we used the C1 platform to investigate transcriptional changes between two models of pulmonary fibrosis – Tgfβ1 overexpression and bleomycin lung injury. In the other experiment, we used Drop-Seq to determine the impact of siRNA mediated FENDRR knockdown. In the C1 experiment we harvested the whole lung of the mouse and generated a single cell suspension whereas in the Drop-Seq experiment, we used cultured normal human lung fibroblasts. Results were subsequently analyzed using unique data pipelines tailored to the specifics of the raw data output from each experiment. Specifically, in Drop-Seq, cell barcodes and UMIs were assigned using one read of the paired-end reads, and the UMIs were used to normalize for asymmetric amplification due to amplification bias. The C1 experiment did not use this internal control, but full transcript matching was possible to enable isoform detection, however, we did not specifically analyze data at the isoform level.

Data quality between C1 and Drop-Seq

In considering the quality of the data between C1 and Drop-Seq the major parameters to consider include the percentage of cells passing threshold, absolute numbers of genes detected, mapping rates, and the biological plausibility of the clustering groups obtained.

As it relates to cells passing thresholds, a greater number of cells were excluded using the C1 filtering steps but this varies depending on the specific experiment. Thus, the variability may be due more to sample preparation or inherent biological variability in the samples rather than technical differences in the protocol.

The Drop-Seq quality filtering pipeline also employs additional filters such as reads from mitochondrial cells that is not used in the C1 pipeline. Including these filters may result in additional cells being discarded. Finally, filtering for doublets or empty barcodes cannot be done in a definitive way using Drop-Seq and thus has to rely on surrogate measurements causing potentially inclusion of doublets or 'empty droplets' or exclusion of true single cell libraries.

Rather surprisingly, the number of genes detected on both platforms after filtering was relatively similar averaging near 3000. This is interesting given that Drop-Seq requires significantly fewer amplification cycles than the C1. This raises the question of whether or not similar amplification cycles would result in fewer genes being detected on the C1. Moreover, we might expect a greater number of total genes to be detected on the C1 given that a total lung cell suspension should demonstrate a greater diversity of genes being expressed than a suspension from a single type of cell. The mapping rate on both platforms ranged from 60-80% and which is consistent with other single cell experiments.

Subpopulations seen in whole mouse lung comparisons

There are several interesting subpopulations that emerge on inspection of the supervised clustering results of the whole mouse lung. In the Tgfβ1 data, there is evidence of EMT as cells that express epithelial markers are also shown to express markers of both Col1a1 and Acta2. There also exists a small subset of cells which expresses epithelial markers, and Col1a1 but not Acta2. This may be indicative of a continuum of differentiation in which cells transition to a final fate through intermediate stages. Another interesting population is the S100a4+ cells which do

not express any fibroblast markers, despite the fact that S100a4 is alternatively known as Fibroblast Specific Protein 1 or FSP-1. Other work has demonstrated that FSP-1 can also be present on inflammatory cells in human and experimental models of liver disease(103). As the Tgfβ1-overexpressing mice are known to exhibit significant inflammatory infiltrate at this stage of treatment, this subset of cells may correspond to the subpopulation of macrophages described in liver injury models. It is notable that Tgfβ1 transcripts were not directly detected in this analysis. This is likely due to the fact that the Tgfβ1 transcript that was expressed on the transgene in these mice is modified in multiple locations and thus the mapping algorithm fails to align the transcript reliably to the reference genome.

Drop-Seq comparison of FENDRR KO vs. control NHLFs

Single cell analysis of FENDRR KO confirms several findings from bulk analysis that are currently in publication by our group. Namely, we confirm on a single cell level that FENDRR ablation results in upregulation of SMA and senescence associated protein markers, namely p21. This modulation of p21 was found in our lab to occur via interaction with PRC2, a histone-modifying complex comprised of multiple proteins and whose functions is modular depending on the specific composition of the complex. This was confirmed based on enrichment of FENDRR RNA with RNA-IP to SUZ12 and EZH2, both components of the PRC2 complex. In this analysis, we also found that several other lncRNAs demonstrated decreased expression relative to the control treated fibroblasts including Xist, Malat1, and Meg3. This may indicate a global effect on lncRNAs that occurs through FENDRR knockdown.

Single cell analysis also confirms previous studies on the incomplete efficiency of RNAi based knockdown of genes(104), and specifically of lncRNA. For lincRNA efficiency between 20-80% are typically noted, depending on the localization of the lncRNA as well as the mechanism of knockdown(105).

Limitations and Future Work

There are several limitations that are important to highlight here. From a technical standpoint, there are limitations inherent to the tools used. For the C1 data, as previously described, there was no internal control to normalize the impact of amplification bias. For the Drop-Seq data, we assumed the isolation of single cells, however it is possible that cell doublets were being created in process of generating the droplets although this situation is unlikely. From a data analysis perspective, our ability to more completely analyze the C1 data was limited due the fact that the bleomycin experiment and the control experiments did not generate gene clusters with useful information. As such, we were not able to create comparisons of, for example, alveolar epithelial cells across the different models. Another important limitation of both platforms and experiments is the transcriptional impact of generating a single cell suspension. In the case of studying pulmonary fibrosis, the ECM and cell-ECM interactions have been shown to play a crucial role in understanding disease pathophysiology. The process of generating a single cell suspension, which involves mechanical digestion, as well as enzymatic digestion of the ECM components that attach cells to each other and the basement membrane, likely triggers important transcriptional changes. Moreover, there is likely a spectrum of vulnerability to cell rupture or apoptosis. This results in enrichment of the

healthiest or most apoptosis resistant cells and thus skews the potential cell populations seen. The process of flowing the cell suspension through the microfluidics of the device may also impart flow-dependent shear forces on the cells that can serve to alter a transcriptional profile. Finally, some portion of the reads in each cell in both experiments are due to ambient RNA which increases the technical noise present in the procedure.

In future work, implementation of some changes to the experimental workflow can lead in better data quality. For example, UMI barcodes have been incorporated to the C1 protocol(106), which may result in better clustering and thus analysis of the downstream data. From a methodological standpoint, to understand the impact of generation of a single cell solution a bulk tissue RNA-seq protocol should be used and the results should be compared to the single cell analysis in order to determine if there are major genes represented in cells that drop out due to preparation of the single cell suspension. Additional, the approximate abundances of the genes can be compared across the two analysis types to see determine if average representation of each gene remains constant. An important breakthrough in single cell genomic analysis will be the development of reagents that are able to ‘freeze’ the transcriptional state of the cell and simultaneously stabilize the cell membrane so that transcriptional changes and cell integrity and simultaneously preserved.

In summary, we have developed and validated two experimental protocols for measuring transcriptional changes in single cells from *in vivo* and *in vitro* models. These protocols will be used in future experiments with clinical samples of patients

with IPF in which we will use both whole tissue as well as cells isolated from these patients.

References

Bibliography

1. Hutchinson J, Fogarty A, Hubbard R, and McKeever T. Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review. *Eur Respir J.* 2015;46(3):795-806.
2. Wilson MS, and Wynn TA. Pulmonary fibrosis: pathogenesis, etiology and regulation. *Mucosal Immunol.* 2009;2(2):103-21.
3. Kolb M, and Collard HR. Staging of idiopathic pulmonary fibrosis: past, present and future. *Eur Respir Rev.* 2014;23(132):220-4.
4. Ley B, Collard HR, and King TE, Jr. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med.* 2011;183(4):431-40.
5. Fulton BG, and Ryerson CJ. Managing comorbidities in idiopathic pulmonary fibrosis. *Int J Gen Med.* 2015;8(309-18.
6. Cavazza A, Rossi G, Carbonelli C, Spaggiari L, Paci M, and Roggeri A. The role of histology in idiopathic pulmonary fibrosis: an update. *Respir Med.* 2010;104 Suppl 1(S11-22.
7. Kuhn C, 3rd, Boldt J, King TE, Jr., Crouch E, Vartio T, and McDonald JA. An immunohistochemical study of architectural remodeling and connective tissue synthesis in pulmonary fibrosis. *Am Rev Respir Dis.* 1989;140(6):1693-703.
8. Leslie KO. Idiopathic pulmonary fibrosis may be a disease of recurrent, tractional injury to the periphery of the aging lung: a unifying hypothesis regarding etiology and pathogenesis. *Arch Pathol Lab Med.* 2012;136(6):591-600.
9. Hecker L, and Thannickal VJ. Nonresolving Fibrotic Disorders: Idiopathic Pulmonary Fibrosis as a Paradigm of Impaired Tissue Regeneration. *Am J Med Sci.* 2011;341(6):431-4.
10. Bringardner BD, Baran CP, Eubank TD, and Marsh CB. The role of inflammation in the pathogenesis of idiopathic pulmonary fibrosis. *Antioxid Redox Signal.* 2008;10(2):287-301.
11. Horowitz JC, and Thannickal VJ. Epithelial-mesenchymal interactions in pulmonary fibrosis. *Semin Respir Crit Care Med.* 2006;27(6):600-12.
12. Moore MW, and Herzog EL. Regulation and Relevance of Myofibroblast Responses in Idiopathic Pulmonary Fibrosis. *Curr Pathobiol Rep.* 2013;1(3):199-208.
13. Hecker L, Logsdon NJ, Kurundkar D, Kurundkar A, Bernard K, Hock T, Meldrum E, Sanders YY, and Thannickal VJ. Reversal of persistent fibrosis in aging by targeting Nox4-Nrf2 redox imbalance. *Sci Transl Med.* 2014;6(231):231ra47.
14. Akhurst RJ, and Hata A. Targeting the TGFbeta signalling pathway in disease. *Nat Rev Drug Discov.* 2012;11(10):790-811.

15. Coker RK, Laurent GJ, Jeffery PK, du Bois RM, Black CM, and McAnulty RJ. Localisation of transforming growth factor beta1 and beta3 mRNA transcripts in normal and fibrotic human lung. *Thorax.* 2001;56(7):549-56.
16. Coker RK, Laurent GJ, Shahzeidi S, Lympany PA, du Bois RM, Jeffery PK, and McAnulty RJ. Transforming growth factors-beta 1, -beta 2, and -beta 3 stimulate fibroblast procollagen production in vitro but are differentially expressed during bleomycin-induced lung fibrosis. *Am J Pathol.* 1997;150(3):981-91.
17. Lee CG, Cho SJ, Kang MJ, Chapoval SP, Lee PJ, Noble PW, Yehualaeshet T, Lu B, Flavell RA, Milbrandt J, et al. Early growth response gene 1-mediated apoptosis is essential for transforming growth factor beta1-induced pulmonary fibrosis. *J Exp Med.* 2004;200(3):377-89.
18. Bonniaud P, Margetts PJ, Kolb M, Schroeder JA, Kapoun AM, Damm D, Murphy A, Chakravarty S, Dugar S, Higgins L, et al. Progressive transforming growth factor beta1-induced lung fibrosis is blocked by an orally active ALK5 kinase inhibitor. *Am J Respir Crit Care Med.* 2005;171(8):889-98.
19. Bonniaud P, Margetts PJ, Ask K, Flanders K, Gauldie J, and Kolb M. TGF-beta and Smad3 signaling link inflammation to chronic fibrogenesis. *J Immunol.* 2005;175(8):5390-5.
20. Derynck R, and Feng XH. TGF-beta receptor signaling. *Biochim Biophys Acta.* 1997;1333(2):F105-50.
21. Rojas A, Padidam M, Cress D, and Grady WM. TGF-beta receptor levels regulate the specificity of signaling pathway activation and biological effects of TGF-beta. *Biochim Biophys Acta.* 2009;1793(7):1165-73.
22. Shi M, Zhu J, Wang R, Chen X, Mi L, Walz T, and Springer TA. Latent TGF-beta structure and activation. *Nature.* 2011;474(7351):343-9.
23. Munger JS, Huang X, Kawakatsu H, Griffiths MJ, Dalton SL, Wu J, Pittet JF, Kaminski N, Garat C, Matthay MA, et al. The integrin alpha v beta 6 binds and activates latent TGF beta 1: a mechanism for regulating pulmonary inflammation and fibrosis. *Cell.* 1999;96(3):319-28.
24. Hakkinen L, Koivisto L, Gardner H, Saarialho-Kere U, Carroll JM, Lakso M, Rauvala H, Laato M, Heino J, and Larjava H. Increased expression of beta6-integrin in skin leads to spontaneous development of chronic wounds. *Am J Pathol.* 2004;164(1):229-42.
25. Reed NI, Jo H, Chen C, Tsujino K, Arnold TD, DeGrado WF, and Sheppard D. The alphavbeta1 integrin plays a critical in vivo role in tissue fibrosis. *Sci Transl Med.* 2015;7(288):288ra79.
26. Puthawala K, Hadjiangelis N, Jacoby SC, Bayongan E, Zhao Z, Yang Z, Devitt ML, Horan GS, Weinreb PH, Lukashev ME, et al. Inhibition of integrin alpha(v)beta6, an activator of latent transforming growth factor-beta, prevents radiation-induced lung fibrosis. *Am J Respir Crit Care Med.* 2008;177(1):82-90.
27. Scotton CJ, and Chambers RC. Molecular targets in pulmonary fibrosis: the myofibroblast in focus. *Chest.* 2007;132(4):1311-21.
28. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, and Quake SR. Reconstructing lineage hierarchies of

the distal lung epithelium using single-cell RNA-seq. *Nature.* 2014;509(7500):371-5.
29. Swiderski RE, Dencoff JE, Floerchinger CS, Shapiro SD, and Hunninghake GW. Differential expression of extracellular matrix remodeling genes in a murine model of bleomycin-induced pulmonary fibrosis. *Am J Pathol.* 1998;152(3):821-8.
30. Crapo JD, Barry BE, Gehr P, Bachofen M, and Weibel ER. Cell number and cell characteristics of the normal human lung. *Am Rev Respir Dis.* 1982;126(2):332-7.
31. Xia H, Bodempudi V, Benyumov A, Hergert P, Tank D, Herrera J, Braziunas J, Larsson O, Parker M, Rossi D, et al. Identification of a cell-of-origin for fibroblasts comprising the fibrotic reticulum in idiopathic pulmonary fibrosis. *Am J Pathol.* 2014;184(5):1369-83.
32. Cool CD, Groshong SD, Rai PR, Henson PM, Stewart JS, and Brown KK. Fibroblast foci are not discrete sites of lung injury or repair: the fibroblast reticulum. *Am J Respir Crit Care Med.* 2006;174(6):654-8.
33. Bertalanffy FD, and Leblond CP. Structure of respiratory tissue. *Lancet.* 1955;269(6905):1365-8.
34. Williams MC. Alveolar type I cells: molecular phenotype and development. *Annu Rev Physiol.* 2003;65(669-95.
35. Hogan BL, Barkauskas CE, Chapman HA, Epstein JA, Jain R, Hsia CC, Niklason L, Calle E, Le A, Randell SH, et al. Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell.* 2014;15(2):123-38.
36. Barkauskas CE, Cronce MJ, Rackley CR, Bowie EJ, Keene DR, Stripp BR, Randell SH, Noble PW, and Hogan BL. Type 2 alveolar cells are stem cells in adult lung. *J Clin Invest.* 2013;123(7):3025-36.
37. Desai TJ, Brownfield DG, and Krasnow MA. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature.* 2014;507(7491):190-4.
38. Basset F, Ferrans VJ, Soler P, Takemura T, Fukuda Y, and Crystal RG. Intraluminal fibrosis in interstitial lung disorders. *Am J Pathol.* 1986;122(3):443-61.
39. Qunn L, Takemura T, Ikushima S, Ando T, Yanagawa T, Akiyama O, Oritsu M, Tanaka N, and Kuroki T. Hyperplastic epithelial foci in honeycomb lesions in idiopathic pulmonary fibrosis. *Virchows Arch.* 2002;441(3):271-8.
40. Gabbiani G. Evolution and clinical implications of the myofibroblast concept. *Cardiovasc Res.* 1998;38(3):545-8.
41. Scheibner KA, Lutz MA, Boodoo S, Fenton MJ, Powell JD, and Horton MR. Hyaluronan fragments act as an endogenous danger signal by engaging TLR2. *J Immunol.* 2006;177(2):1272-81.
42. Raghu G, Striker LJ, Hudson LD, and Striker GE. Extracellular matrix in normal and fibrotic human lungs. *Am Rev Respir Dis.* 1985;131(2):281-9.
43. Xu Y, Mizuno T, Sridharan A, Du Y, Guo M, Tang J, Wikenheiser-Brokamp KA, Perl AT, Funari VA, Gokey JJ, et al. Single-cell RNA sequencing

identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI Insight.* 2016;1(20):e90558.
44. Rennard SI, Stier LE, and Crystal RG. Intracellular degradation of newly synthesized collagen. *J Invest Dermatol.* 1982;79 Suppl 1(77s-82s.
45. Laurent GJ. Rates of collagen synthesis in lung, skin and muscle obtained in vivo by a simplified method using [3H]proline. *Biochem J.* 1982;206(3):535-44.
46. Broekelmann TJ, Limper AH, Colby TV, and McDonald JA. Transforming growth factor beta 1 is present at sites of extracellular matrix gene expression in human pulmonary fibrosis. *Proc Natl Acad Sci U S A.* 1991;88(15):6642-6.
47. McDonald JA, Broekelmann TJ, Matheke ML, Crouch E, Koo M, and Kuhn C, 3rd. A monoclonal antibody to the carboxyterminal domain of procollagen type I visualizes collagen-synthesizing fibroblasts. Detection of an altered fibroblast phenotype in lungs of patients with pulmonary fibrosis. *J Clin Invest.* 1986;78(5):1237-44.
48. Raghu G, Chen YY, Rusch V, and Rabinovitch PS. Differential proliferation of fibroblasts cultured from normal and fibrotic human lungs. *Am Rev Respir Dis.* 1988;138(3):703-8.
49. Yanai H, Shteinberg A, Porat Z, Budovsky A, Braiman A, Ziesche R, and Fraifeld VE. Cellular senescence-like features of lung fibroblasts derived from idiopathic pulmonary fibrosis patients. *Aging (Albany NY).* 2015;7(9):664-72.
50. Desmouliere A, Geinoz A, Gabbiani F, and Gabbiani G. Transforming growth factor-beta 1 induces alpha-smooth muscle actin expression in granulation tissue myofibroblasts and in quiescent and growing cultured fibroblasts. *J Cell Biol.* 1993;122(1):103-11.
51. Zhang K, Rekhter MD, Gordon D, and Phan SH. Myofibroblasts and their role in lung collagen gene expression during pulmonary fibrosis. A combined immunohistochemical and in situ hybridization study. *Am J Pathol.* 1994;145(1):114-25.
52. Willis BC, duBois RM, and Borok Z. Epithelial origin of myofibroblasts during fibrosis in the lung. *Proc Am Thorac Soc.* 2006;3(4):377-82.
53. Kalluri R, and Neilson EG. Epithelial-mesenchymal transition and its implications for fibrosis. *J Clin Invest.* 2003;112(12):1776-84.
54. Pandit KV, Milosevic J, and Kaminski N. MicroRNAs in idiopathic pulmonary fibrosis. *Transl Res.* 2011;157(4):191-9.
55. Moeller A, Ask K, Warburton D, Gauldie J, and Kolb M. The bleomycin animal model: a useful tool to investigate treatment options for idiopathic pulmonary fibrosis? *Int J Biochem Cell Biol.* 2008;40(3):362-82.
56. Burns AR, Smith CW, and Walker DC. Unique structural features that influence neutrophil emigration into the lung. *Physiol Rev.* 2003;83(2):309-36.
57. Doherty DE, Hirose N, Zagarella L, and Cherniack RM. Prolonged monocyte accumulation in the lung during bleomycin-induced pulmonary fibrosis. A noninvasive assessment of monocyte kinetics by scintigraphy. *Lab Invest.* 1992;66(2):231-42.
58. Haslett C, Shen AS, Feldsien DC, Allen D, Henson PM, and Cherniack RM. 111Indium-labeled neutrophil migration into the lungs of bleomycin-treated

rabbits assessed noninvasively by external scintigraphy. *Am Rev Respir Dis.* 1989;140(3):756-63.
59. Adler KB, Low RB, Leslie KO, Mitchell J, and Evans JN. Contractile cells in normal and fibrotic lung. *Lab Invest.* 1989;60(4):473-85.
60. Thannickal VJ, and Horowitz JC. Evolving concepts of apoptosis in idiopathic pulmonary fibrosis. *Proc Am Thorac Soc.* 2006;3(4):350-6.
61. Zhang HY, Gharaee-Kermani M, Zhang K, Karmiol S, and Phan SH. Lung fibroblast alpha-smooth muscle actin expression and contractile phenotype in bleomycin-induced pulmonary fibrosis. *Am J Pathol.* 1996;148(2):527-37.
62. Moore BB, Kolodsick JE, Thannickal VJ, Cooke K, Moore TA, Hogaboam C, Wilke CA, and Toews GB. CCR2-mediated recruitment of fibrocytes to the alveolar space after fibrotic injury. *Am J Pathol.* 2005;166(3):675-84.
63. Korfhagen TR, Swantz RJ, Wert SE, McCarty JM, Kerlakian CB, Glasser SW, and Whitsett JA. Respiratory epithelial cell expression of human transforming growth factor-alpha induces lung fibrosis in transgenic mice. *J Clin Invest.* 1994;93(4):1691-9.
64. Christensen PJ, Goodman RE, Pastoriza L, Moore B, and Toews GB. Induction of lung fibrosis in the mouse by intratracheal instillation of fluorescein isothiocyanate is not T-cell-dependent. *Am J Pathol.* 1999;155(5):1773-9.
65. Fisher CE, Ahmad SA, Fitch PM, Lamb JR, and Howie SE. FITC-induced murine pulmonary inflammation: CC10 up-regulation and concurrent Shh expression. *Cell Biol Int.* 2005;29(10):868-76.
66. Chiang CS, Liu WC, Jung SM, Chen FH, Wu CR, McBride WH, Lee CC, and Hong JH. Compartmental responses after thoracic irradiation of mice: strain differences. *Int J Radiat Oncol Biol Phys.* 2005;62(3):862-71.
67. Haston CK, and Travis EL. Murine susceptibility to radiation-induced pulmonary fibrosis is influenced by a genetic factor implicated in susceptibility to bleomycin-induced pulmonary fibrosis. *Cancer Res.* 1997;57(23):5286-91.
68. Davis GS, Leslie KO, and Hemenway DR. Silicosis in mice: effects of dose, time, and genetic strain. *J Environ Pathol Toxicol Oncol.* 1998;17(2):81-97.
69. Lakatos HF, Burgess HA, Thatcher TH, Redonnet MR, Hernady E, Williams JP, and Sime PJ. Oropharyngeal aspiration of a silica suspension produces a superior model of silicosis in the mouse when compared to intratracheal instillation. *Exp Lung Res.* 2006;32(5):181-99.
70. Moore BB, and Hogaboam CM. Murine models of pulmonary fibrosis. *Am J Physiol Lung Cell Mol Physiol.* 2008;294(2):L152-60.
71. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, and Cui Q. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 2013;41(Database issue):D983-6.
72. Kim S, Herazo-Maya JD, Kang DD, Juan-Guardela BM, Tedrow J, Martinez FJ, Sciurba FC, Tseng GC, and Kaminski N. Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genomics.* 2015;16(924.

73. Chandler H, Patel H, Palermo R, Brookes S, Matthews N, and Peters G. Role of polycomb group proteins in the DNA damage response--a reassessment. *PLoS One.* 2014;9(7):e102968.
74. . In: Corporation F ed.; 2016.
75. . Fluidigm | Script Hub. https://www.fluidigm.com/c1openapp/scripthub. 3/16/2017.
76. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, and Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods.* 2013;10(11):1096-8.
77. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell.* 2015;161(5):1202-14.
78. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P, and Linnarsson S. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods.* 2014;11(2):163-6.
79. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, and Enard W. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell.* 2017;65(4):631-43 e4.
80. Fluidigm. 2016.
81. Dabney J, and Meyer M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques.* 2012;52(2):87-94.
82. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, and Taipale J. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods.* 2011;9(1):72-4.
83. Picelli S, Faridani OR, Bjorklund AK, Winberg G, Sagasser S, and Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc.* 2014;9(1):171-81.
84. Bray NL, Pimentel H, Melsted P, and Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* 2016;34(5):525-7.
85. McCarthy D WQaCK. https://github.com/davismcc/scater; 2016.
86. Langfelder P, Zhang B, and Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics.* 2008;24(5):719-20.
87. Nemesh J. 2016:Tools for converting raw reads to DGE matrix.
88. Nemesh J. Drop-seq Core Computational Protocol. 2016.
89. Institute B. http://broadinstitute.github.io/picard., 2016.
90. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21.
91. Nemesh J. 2016.
92. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, Garcia Giron C, Hourlier T, et al. The Ensembl gene annotation system. *Database (Oxford).* 2016;2016(

93. Satija R. http://www.satijalab.org/seurat; 2015.
94. Amir el AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, and Pe'er D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol.* 2013;31(6):545-52.
95. van der Maaten L. Visualizing Data using t-SNE. *J Mach Learn.* 2008;Res. 9(2579–605.
96. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res.* 2014;42(Database issue):D472-7.
97. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, D'Eustachio P, and Stein L. Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers (Basel).* 2012;4(4):1180-211.
98. Akamatsu T, Arai Y, Kosugi I, Kawasaki H, Meguro S, Sakao M, Shibata K, Suda T, Chida K, and Iwashita T. Direct isolation of myofibroblasts and fibroblasts from bleomycin-injured lungs reveals their functional similarities and differences. *Fibrogenesis Tissue Repair.* 2013;6(1):15.
99. Perucca P, Cazzalini O, Madine M, Savio M, Laskey RA, Vannini V, Prosperi E, and Stivala LA. Loss of p21 CDKN1A impairs entry to quiescence and activates a DNA damage response in normal fibroblasts induced to quiescence. *Cell Cycle.* 2009;8(1):105-14.
100. Ishige T, Nishimura M, Satoh M, Fujimoto M, Fukuyo M, Semba T, Kado S, Tsuchida S, Sawai S, Matsushita K, et al. Combined Secretomics and Transcriptomics Revealed Cancer-Derived GDF15 is Involved in Diffuse-Type Gastric Cancer Progression and Fibroblast Activation. *Sci Rep.* 2016;6(21681.
101. Kalluri R, and Zeisberg M. Fibroblasts in cancer. *Nat Rev Cancer.* 2006;6(5):392-401.
102. Sureshbabu A, Okajima H, Yamanaka D, Shastri S, Tonner E, Rae C, Szymanowska M, Shand JH, Takahashi S, Beattie J, et al. IGFBP-5 induces epithelial and fibroblast responses consistent with the fibrotic response. *Biochem Soc Trans.* 2009;37(Pt 4):882-5.
103. Osterreicher CH, Penz-Osterreicher M, Grivennikov SI, Guma M, Koltsova EK, Datz C, Sasik R, Hardiman G, Karin M, and Brenner DA. Fibroblast-specific protein 1 identifies an inflammatory subpopulation of macrophages in the liver. *Proc Natl Acad Sci U S A.* 2011;108(1):308-13.
104. Leonhardt C, Schwake G, Stogbauer TR, Rappl S, Kuhr JT, Ligon TS, and Radler JO. Single-cell mRNA transfection studies: delivery, kinetics and statistics by numbers. *Nanomedicine.* 2014;10(4):679-88.
105. Lennox KA, and Behlke MA. Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. *Nucleic Acids Res.* 2016;44(2):863-77.
106. T.Kouno SK, M.Mendez, I.Abugessaisa, J.Shin and C.Plessy. 2016.